ICASE

A COMPARISON OF LABORATORY EXPERIMENTS WITH A MODEL EQUATION FOR WATER WAVES

J. L. Bona
W. G. Pritchard
and
L. R. Scott

Report No. 80-26
October 16, 1980

(NASA-CR-185778) A COMPARISON OF LABORATORY EXPERIMENTS WITH A MODEL EQUATION FOR WATER WAVES (ICASE) 110 p

N89-71318

Unclas
00/34 0224297

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia

Operated by the

UNIVERSITIES SPACE USRA RESEARCH ASSOCIATION

CONTENTS

A COMPARISON OF LABORATORY

EXPERIMENTS WITH A MODEL EQUATION FOR WATER WAVES

J. L. Bona*
University of Chicago

W. G. Pritchard†
University of Essex, England

and

L. R. Scott†§
University of Michigan

ABSTRACT

The aim of this paper is to assess how well the equation

$$\eta_t + \eta_x + \tfrac{3}{2}\eta\eta_x - \mu\eta_{xx} - \tfrac{1}{6}\eta_{xxt} = 0$$

describes the propagation of water waves in a laboratory experiment. Here x is the horizontal coordinate, t is the time, η is the displacement of the surface of the water from its equilibrium position and μ is a real constant.

A numerical scheme has been developed to solve the above equation for $x, t > 0$, subject to the initial condition $\eta(x,0) = 0$ and to the boundary condition $\eta(0,t) = h(t)$, where h is a specified function. In the present context, h can be thought of as an amplitude at one end of a long channel. The numerical scheme that has been used is an explicit, unconditionally stable scheme having fourth-order accuracy in both space and time. A rigorous analysis of the errors inherent in the numerical scheme, as well as convergence tests of the code, are presented.

Quantitative comparisons between the model and our laboratory experiments typically showed differences of around 8%, increasing to about 30% at the larger wave amplitudes used in the experiments. The agreement was found to be considerably worse than this if damping of the waves was not included (i.e., if $\mu = 0$). An interpretation of the experimental results in terms of the model equation is given and attempts are made to assess some of the factors leading to the observed differences at the larger amplitudes.

*Work was supported by the Science Research Council (U.K.).

†The second and third authors were supported under NASA Contract No. NAS1-14101 while they were in residence at the Institute of Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA 23665. A part of the research was also carried out at the Brookhaven National Laboratory.

§Work was also supported under AFOSR Contract No. F49620-79-C-0149.

1. INTRODUCTION

This study attempts to assess a particular model for the unidirectional propagation of water waves as a predictor of the results of a set of laboratory experiments. Although the first one-dimensional model for the propagation of weakly nonlinear waves in shallow water was proposed last century (Korteweg & de Vries 1895) it is only in recent years that any serious attempts have been made to test this model in practice. The main reasons for the long delay, with regard to this and similar models, derive from the difficulty in obtaining solutions to the equations, other than the rather special solitary-wave and cnoidal-wave solutions. However, it is now feasible to devise sound numerical schemes to integrate some of the model equations: we shall propose one such scheme, for a particular model, and use it to test how well the model describes the experimental situation.

To a certain extent such a programme has been carried out by Zabusky & Galvin (1971) and by Hammack & Segur (1974). Both these studies suggested that the Korteweg-de Vries equation (henceforth to be referred to as the KdV equation) gave a reasonably good qualitative account of the experiments. But the quantitative agreement was not striking, mainly because the studies were made under conditions to which the model should not necessarily be expected to apply and because the comparison procedures involving an approximate transformation of the initial data can lead to significant errors (cf. appendix A).

An important part of the present work is the numerical integration of the model equation under scrutiny. Since the interpretation of laboratory experiments inevitably gives rise to many difficulties, it seemed appropriate to be absolutely sure that the numerical solutions were close approximations to the solutions of the model equations. Thus, in § 3 we give a detailed account of the numerical schemes employed, together with rigorous estimates of the error bounds and convergence tests of the scheme.

The structure of the paper is as follows. In § 2 we first discuss model equations apposite to the present study, together with the concomitant assumptions built in during the modelling, and then examine the more important empirical design criteria. Some theoretical properties of the solutions to the model to be tested are given in §3. These are needed for the error estimates for the numerical scheme which is described and analyzed in §§ 4,5. Convergence tests for the scheme are also given. In §6 the experimental procedure is described and in §7 the main results are presented. At the smaller amplitudes used in the experiments the model equation appears to have given a fairly good description of the experimental results, but at larger amplitudes the model did not work so well. Some possible sources for these discrepancies are examined in §§ 7.4, 7.5. A résumé of the main results is given in § 8.

2. EXPERIMENTAL DESIGN

2.1 Model equations

Consider two-dimensional surface waves propagating along a uniform horizontal channel. Suppose that the waves propagate only in the positive x-direction and that the undisturbed depth of the liquid in the channel is d. All the variables used here are dimensionless, with the length scale taken to be the equilibrium depth, d, and the time scale to be $(d/g)^{\frac{1}{2}}$, g being the acceleration due to gravity. Let t be the time and let $\eta = \eta(x,t)$ represent the vertical displacement of the surface of the liquid from its equilibrium position. The horizontal coordinate x is measured along the channel. If the horizontal scale, δ^{-1}, of the motions is large and the maximum amplitude ε of the waves is sufficiently small, then a model for the propagation of irrotational waves is afforded by the KdV equation (see Whitham 1974)

$$\eta_t + \eta_x + \tfrac{3}{2}\eta\eta_x + \tfrac{1}{6}\eta_{xxx} = 0 . \qquad \text{(KdV)}$$

The primary terms η_t, η_x represent a uniform translation of a wave and it is proposed that the secondary terms account respectively for the modification of the wave through the separate influences of nonlinear and dispersive effects. The relative importance of the nonlinear and the dispersive effects is given by the parameter $S = \varepsilon\delta^{-2}$, an important assumption in the derivation of KdV being that this parameter is O(1) (cf. Meyer 1972, Whitham 1974).† These considerations suggest the introduction of a new dependent variable N and new independent variables ξ, τ such that

$$\eta = \varepsilon N , \quad x = \varepsilon^{-1/2}\xi , \quad t = \varepsilon^{-1/2}\tau .$$

Thus, by assumption, N and its derivatives with respect to the new independent variables are all O(1), and it follows that (KdV) can be written as

$$N_\tau + N_\xi + \tfrac{3}{2}\varepsilon N N_\xi + \tfrac{1}{6}\varepsilon N_{\xi\xi\xi} = O(\varepsilon^2) , \qquad (2.2)$$

showing explicitly the relative sizes of the various terms. (On the

† Here, and in what follows, the symbol O(·) will be used informally in the way that is common in the construction and formal analysis of model equations for physical phenomena. A strict usage could be maintained in which the relevant limit is $\varepsilon \downarrow 0$, $\delta \downarrow 0$, S constant.

For waves of wavelength λ and amplitude a, $S = a\lambda^2/d^3$.

right-hand-side of (2.2) we have indicated the relative size of the terms neglected in the formal derivation of the KdV model). A physical interpretation of (2.2) is that the small nonlinear and dispersive corrections can accumulate and, on time scales τ of $O(\varepsilon^{-1})$ (or $t = O(\varepsilon^{-3/2}))$, have made important modifications to the initial waveform. Moreoever, since the terms neglected in (2.2) are $O(\varepsilon^2)$, it follows that, on time scales $\tau = O(\varepsilon^{-2})$ (or $t = O(\varepsilon^{-5/2}))$, the model can no longer be formally justified.

Because of the orders of magnitude of the terms in (2.2) an alternative model for the same physical situation, valid to the same accuracy as the KdV equation, is the equation (see Peregrine 1966, Benjamin et al, 1972)

$$N_\tau + N_\xi + \tfrac{3}{2}\varepsilon N N_\xi - \tfrac{1}{6} N_{\xi\xi\tau} = O(\varepsilon^2) .$$

In terms of the physical variables $\eta(x,t)$ this model takes the form

$$\eta_t + \eta_x + \tfrac{3}{2}\eta\eta_x - \tfrac{1}{6}\eta_{xxt} = 0 . \qquad \text{(M)}$$

Thus, in summary, (KdV) and (M) have been proposed as models for the propagation of water waves under the following conditions.

(i) The waves effectively propagate in one direction. This precludes the possibility of interactions with reflected waves and, in particular, it means that any variations in the depth of the channel should occur on length scales much larger than the horizontal scale of the waves.

(ii) The wave amplitudes are small (i.e. $\varepsilon \ll 1$) and the horizontal length scale of the waves is large (i.e. $\delta \ll 1$).

(iii) The nonlinear and dispersive effects are comparable: $\varepsilon\delta^{-2} = O(1)$.

(iv) The waves arise on an irrotational flow.

(v) There is no mechanical degradation of energy.

(vi) The influence of surface tension is negligible (though this restriction can be relaxed, cf. Korteweg & de Vries 1895).

We can expect significant modifications to a waveform on a time scale $O(\varepsilon^{-3/2})$ and, from a formal viewpoint, the model can not be justified on times which are $O(\varepsilon^{-5/2})$.

2.2 Previous studies

In 1971 Zabusky & Galvin reported some experiments in which a train of initially sinusoidal waves propagated into still water. At stations further along the channel they found that, after the first couple of wave crests had passed, the wave profiles were very nearly periodic in time. This property suggested a numerical experiment in which a periodic version of KdV was integrated, using a sinusoidal waveform as the initial data. Then, to compare the numerical computations with the experiments, the long-wave speed for linear disturbances was used to provide a kind of equivalence between time in the periodic problem and position in the experimental configuration. The experiments were made at values of S equal to 22, 482 and 777. Fairly good qualitative agreement was obtained between the predicted wave shapes and those observed experimentally, but quantitative comparisons were not made, principally because viscous effects had a significant influence on the experimental results.

A study similar in concept to the programme to be described here was made by Hammack (1973). Water was displaced at one end of a channel generating an isolated waveform, the passage of which was observed at various positions along the channel. Comparisons made between the observed profiles and numerical solutions of (M) showed good qualitative agreement but, since the computations were not very accurate and since viscous effeots were again important, detailed quantitative comparisons were not made. For these experiments the value of S lay between about 1 and 10.

In a subsequent experiment Hammack & Segur (1974) also followed the evolution of an isolated waveform propagating along a channel. Using the inverse-scattering methods developed for the KdV equation they predicted both the number of solitons to emerge from the initial waveform and also the amplitude of the largest soliton. The predicted number of emergent solitons was in agreement with the experimental observations, but the predicted amplitude of the leading soliton (after making a correction for viscous damping along the lines suggested by Keulegan 1948) differed by about 15-20% from the observed values. These experiments were carried out at values of S ranging between 50 and 600.

In each of the above studies, the theoretical solutions were obtained from the solution of a pure initial-value problem. However, the initial data set was not obtained in the form required for the theoretical model, necessitating a transformation of the data. Because the transformation employed was inexact this may have led to significant errors in the solution (see appendix A).

2.3 Allowing for dissipation

One of the main conclusions to be drawn from the previous experimental studies is that useful quantitative predictions can be made only by taking account of dissipative effects. On the scale of the present experiment the main sources of wave damping appear to derive from viscous dissipation in the boundary layers on the sides and bottom of the channel, from the influence of the meniscus at the side walls of the channel and perhaps from damping at the free surface (see Barnard, Mahony & Pritchard 1977 and Mahony & Pritchard 1980). It is possible to incorporate the effects of the boundary layers on the walls of the channel into the theories described above (see Kakutani & Matsuuchi 1975), but there are both empirical and theoretical uncertaintities about the representation of the effects at the free surface and at the meniscus (see Miles 1967, Mei & Liu 1973). Thus, any attempts to account for dissipation must, to a certain extent, be guesswork.

The rationale behind the construction of models such as those described in § 2.1 is that the various corrections to the primary waveform can be calculated independently, with a composite model formed by including the modifications additively (on the assumption that the coupling between them is negligible). Because of this it is sufficient, for the time being, to consider the effects of damping only on waves of extremely small amplitude, so that a linear model is applicable. Then the dispersion relation between the frequency ω and the wavenumber k is given by

$$\omega = k(1-\tfrac{1}{6}k^2) \quad (\mathrm{KdV}), \quad \omega = k(1+\tfrac{1}{6}k^2)^{-1} \quad (\mathrm{M}), \quad \omega = (k \tanh k)^{1/2} \quad (\mathrm{exact}). \qquad (2.4)$$

By construction the phase speeds ω/k for each of these relations are different only at the fourth order in k.

The theory of Kakutani & Matsuuchi (1975) indicates that the effect of dissipation in the boundary layers on the rigid surfaces of the channel is comparable with the nonlinear and the dispersive corrections from the inviscid theories when the wavenumber k is $O(R^{-1/5})$. Here the Reynolds number $R=(gd^3)^{1/2}/\nu$, where ν represents the kinematic viscosity of the fluid. Under these conditions Kakutani & Matsuuchi show that the dispersion relation for (KdV) should be modified to

$$\omega = k\left(1 - \tfrac{1}{6}k^2\right) - i\rho\,|k|^{1/2}, \tag{2.5}$$

where ρ is a complex number depending on R . Thus, not only do the boundary layers induce a damping of the waves but they also affect the phase speed slightly. Moreover, the analysis indicates that the boundary-layer damping can be neglected only when $(kR)^{-1/2} \ll k^2$. So, as a rough guide, we cannot expect to be able to discard dissipative effects when the water depth is less than a metre. The kind of damping introduced in (2.5) can, of course, also be incorporated into model (M). A term of this kind introduces a pseudo-differential operator in each of the model equations.

However, as indicated above, the boundary-layer theory considerably underestimates the damping rate (by about 40% on the scale of the present experiment, according to Mahony & Pritchard 1980). Because of the inadequacy of the theory in this respect we decided to use an *ad hoc* representation of the wave damping to preserve the simple structure of the model equation, rather than attempting a more complicated representation that could not be totally justified anyway. Thus, we shall suppose that a wave of wavenumber k is damped at a rate proportional to k^2 , having the effect of introducing a term $\mu\eta_{xx}$, $\mu \in \mathbb{R}$, into the model equation, and this can easily be incorporated into the numerical scheme of § 3.

For the experiments to be described the waves were generated by a forced motion at a frequency ω_0 , with the result that most of the energy should have resided in a single wavenumber k_0 , say. Then, by choosing μ such that the damping of waves of wavenumber k_0 agreed with the experimental decay rate at very small amplitudes, we should at least have modelled correctly the dissipation of the fundamental, even if other wavenumbers are likely to have been dissipated at an incorrect rate. (This statement is of course based on the presumption that the wave damping depended linearly on η, which might not be justified in the case of damping derived from the effects of the menisci).

These very considerations indicate that we need to be circumspect about the representation of the dissipative effects, a point we shall consider in more detail in § 5.4. However, for the present let us take the model equation in the form

$$\eta_t + \eta_x + \tfrac{3}{2}\eta\eta_x - \mu\eta_{xx} - \tfrac{1}{6}\eta_{xxt} = 0 . \qquad \begin{cases}(2.6),\\(M^*)\end{cases}$$

2.4 Mathematical considerations

Three kinds of mathematical problems have been studied in connection with (KdV) or (M).

(i) Pure initial-value problems. For this class of problem it is supposed that the surface profile is known at some instant, say $t=0$. Mathematically this amounts to the specification

$$\eta(x,0) = g(x) , \quad \text{for} \quad x \in \mathbb{R} . \qquad (2.7)$$

Interest is focussed on the solution of (KdV) or (M), defined for $t \geq 0$, which agrees with g at $t=0$. If g is an element of a function class comprised of smooth functions that decay to zero sufficiently rapidly at ±∞, then it is known that the specification (2.7) constitutes a well-posed problem in conjunction with (M)(eg. see Benjamin et al. 1972) or in conjunction with (KdV) (eg. see Bona & Smith 1975).

A physical realization of this formulation of the problem can be achieved in a long channel by establishing a wavetrain of restricted spatial extent that propagates from one end of the channel to the other. A photograph of the water surface at some instant could be used to determine the initial datum g and the wave profile at later times could be compared with, say, numerical solutions to the model problem. (This, in essence, is the kind of programme carried out by Hammack & Segur 1974. However, in their case, the determination of g(x) was made from a temporal wave record $g(x_0,t), x_0$ fixed, together with the leading order approximation $\eta_t + \eta_x = 0$ for the wave field. It is shown in appendix A that such a procedure can lead to significant errors and should be avoided.)

(ii) Periodic initial-value problems. These problems are the same as described in (i) except that the initial datum, g, is given a periodic function. Again, the mathematical problems for (KdV) and for (M) are well posed. However, the physical realization of such a model is very difficult to achieve. (Zabusky & Galvin 1971 used numerical solutions to a problem of this kind to explain qualitatively the behaviour of waves generated by the periodic motion of a wavemaker at one end of the channel, cf. § 2.2.)

(iii) Initial- and boundary-value problems. For this class of problem we are interested in solutions $\eta(x,t)$ for $x, t > 0$, to the model equations, subject to the conditions

$$\eta(x,0) = g(x)\ , \ x \geqslant 0\ , \ \text{and} \quad \eta(0,t) = h(t), \ t \geqslant 0. \tag{2.8}$$

For consistency we suppose that $g(0) = h(0)$. It has been shown by Bona & Bryant (1973) that, under these conditions, (M) constitutes a well-posed problem if g,h are suitably smooth functions. However, a complete theory for (KdV) has not yet appeared.

In physical terms g represents the initial configuration of the water surface; usually we would expect at the outset that the water is undisturbed, in which case we would have $g = 0$. The function h(t) represents an imposed amplitude, of the water surface at the left-hand end of the channel. Thus, we might think of this kind of problem as a model for waves with known amplitude initiated at one end of a long channel.

2.5 Practical considerations. The issues raised in the preceding discussion impose considerable restrictions on the experimental design. But, in addition, if the models are to be of any real practical value they should be applicable to the kind of situation that usually obtains in the laboratory, namely the propagation of waves arising from the forcing effects of a wavemaker at one end of a channel. Since wavemakers are usually driven in a periodic motion, it would be nice to preserve this feature as well. Indeed, such forcing would be desirable here because the imposed frequency effectively establishes a length scale for the motions, allowing a fairly precise specification of the parameter S. In order to meet these requirements and to simplify the experimental procedure, it would appear that the most suitable kind of mathematical problem to model is the initial- and boundary-value formulation. (One of the main empirical difficulties in modelling the pure initial-value formulation is that of obtaining an instantaneous spatial measurement of the wave field. But also, in our case, the wave tank available would not have been long enough for such an experiment.) A convenient experimental procedure would be to start with the channel free of motion and then to set the wavemaker working at a fixed frequency and amplitude. This would initiate a train of waves that would propagate along the channel, retaining their unidirectional quality until they reach the end of the channel, at which point the experiment would have to cease. The boundary condition h(t) in (2.8) could be specified by a temporal record of the wave amplitude (taken at a position far enough away from the wavemaker to avoid confusing the free waves with the parasitic field localized near the paddle).

The wave tank available in our laboratory was only 5½m long. So, in order to allow enough time for the waves to show significant modifications before reaching the end of the tank, the basic wavelength had not to be too large. On the other hand, it had to be larger than the channel width (30 cm) to avoid spontaneous generation of transverse modes. A reasonable compromise for the wavelength appeared to be 36 cm. We decided to use a wavelength-to-depth ratio of 12:1.

In principle we would like the experiment to cover a range of wave amplitudes for which the parameter S, measuring the relative importance of nonlinear and dispersive effects, spans a fairly representative range of parameter space. Under the above conditions, S would take a value of 0.1 at a wave amplitude of 0.002 cm and would be 10 at a wave amplitude of 0.2 cm. So, to be sure of achieving linear motions at one end of the parameter range, it would be necessary to use very small wave amplitudes. Fortunately, in our experiments, this did not pose any major difficulties.

3. PROPERTIES OF THE EXACT SOLUTION OF THE MODEL EQUATION

In this section we study properties of the solution of the initial- and boundary-value problem

$$\left.\begin{array}{l} \eta_t + \alpha\eta_x + \beta\eta\eta_x - \mu\eta_{xx} - \gamma\eta_{xxt} = 0\,, \\ \eta(x,0) = 0\,, \quad \eta(0,t) = h(t)\,, \end{array}\right\} \text{for } x,t \geqslant 0\,, \tag{3.1}$$

where α, β, μ are non-negative constants and γ is a positive constant. We shall first discuss the questions of existence, uniqueness and a priori boundedness of η. Then, in preparation for a posteriori error estimates to be derived in §5, bounds for derivatives of η are given in terms of assumed bounds on η. Finally, it is shown that η decays exponentially in space, justifying the truncation of the spatial domain in numerical calculations.

3.1. Existence, uniqueness and a priori bounds for η

Suppose that the boundary data is 'smooth' in the sense that, for a given T > 0 and an integer $\ell \geqslant 1$,

$$h \in \mathcal{C}^{\ell}([0,T]) \quad \text{and} \quad h(0) = 0\,. \tag{3.2}$$

Then, using the techniques of Bona & Bryant (1972), it follows that (3.1) has a unique solution $\eta \in \mathcal{C}_T^{\ell,k}$; that is, $\left(\frac{\partial}{\partial x}\right)^i \left(\frac{\partial}{\partial t}\right)^j \eta(x,t)$ exists and is continuous on $[0,\infty[\times [0,T]$, for i = 0,1,..., ℓ and j = 0,1,...,k. (Here k may be any positive integer). Furthermore, these derivatives of η all tend to zero as $x \to \infty$, and η, η_x are square integrable in x on $[0,\infty]$. If $|h(t)|$ and $|h'(t)|$ are bounded by some constant, say M, for $t \in [0,T]$, then using the methods of Bona & Bryant (1972) it can be shown, for $t \in [0,T]$,that

$$\max\{|\eta(x,t)| : x \geqslant 0\,,\ t \in [0,T]\} \leqslant b_1 e^{b_2 t}\,, \tag{3.3}$$

where b_1 , b_2 are constants depending only on $\alpha, \beta, \mu, \gamma$ and M . In addition it follows that the solution to (3.1) satisfies the equation (cf. Bona & Bryant 1972)

$$\left.\begin{aligned} \eta_t(x,t) = h'(t)e^{-x/\sqrt{\gamma}} &+ \int_0^\infty \tilde{K}(x,y)(\alpha\eta + \tfrac{1}{2}\beta\eta^2)(y,t)\,dy \\ &+ \frac{\mu}{\gamma}\left[h(t)e^{-x/\sqrt{\gamma}} - \eta(x,t)\right] - \mu\int_0^\infty \tilde{H}(x,y)\,\eta(y,t)\,dy \ , \end{aligned}\right\} \quad (3.4)$$

where

$$\left.\begin{aligned} &\tilde{K}(x,y)\ (\equiv K_A(x,y) + K(x,y)) = \frac{1}{2\gamma}\left[e^{-(x+y)/\sqrt{\gamma}} + \mathrm{sgn}\,(x-y)\,e^{-|x-y|/\sqrt{\gamma}}\right] \\ \text{and}\quad &\tilde{H}(x,y)\ (\equiv H_A(x,y) + H(x,y)) = \frac{1}{2\gamma^{3/2}}\left[e^{-(x+y)/\sqrt{\gamma}} - e^{-|x-y|/\sqrt{\gamma}}\right]. \end{aligned}\right\} \quad (3.5)$$

The numerical scheme to be described in §4 is based on this formulation.

From the definitions it follows, for any non-negative integer k, that

$$\left.\begin{aligned} &\max\left\{\int_0^x \left|\left(\frac{\partial}{\partial y}\right)^k \tilde{K}(x,y)\right| dy\ , \int_x^\infty \left|\left(\frac{\partial}{\partial y}\right)^k \tilde{K}(x,y)\right| dy\ :\ x > 0\right\} \le \gamma^{-(k+1)/2}, \\ &\max\left\{\int_0^x \left|\left(\frac{\partial}{\partial y}\right)^k \tilde{H}(x,y)\right| dy, \int_x^\infty \left|\left(\frac{\partial}{\partial y}\right)^k \tilde{H}(x,y)\right| dy : x > 0\right\} \le \gamma^{-(k+2)/2}. \end{aligned}\right\} \quad (3.6)$$

Remarks. (i) The a priori bound (3.3) can be improved considerably. For example, in some separate work we have been able to show that $\max\{|\eta|\}$ grows no faster than t. However, (3.3) is sufficient to obtain a posteriori estimates (see §5) which show that there is essentially no growth in the maximum of $|\eta|$, provided the same holds true for a discrete approximation to η .

(ii) The above theory holds when (3.1) is posed with non-zero initial data $\eta(x,0) = g(x)$, provided that $g \in \mathscr{C}^k([0,\infty])$ for $k \geq 2$, that g and its derivatives tend to zero as $x \to \infty$, that g, g' are square integrable and that $g(0) = h(0)$.

3.2. Bounds for the derivatives of η

Bounds on the temporal and spatial derivatives of η are to be derived in terms of assumed bounds on the maximum of $|\eta|$ itself. Thus for T > 0, define

$$\sigma(T) = \max\left\{|\eta(x,t)| : x \geqslant 0,\ t \in [0,T]\right\}. \tag{3.7}$$

We shall use the notation

$$\|\phi\| = \max\left\{|\phi(x)| : x \geqslant 0\right\}, \tag{3.8}$$

where ϕ represents η or its derivatives.

Bounds for η_t can be obtained directly from (3.4) through an application of Hölder's inequality (together with (3.6)). These imply that

$$\|\eta_t(\cdot,t)\| \leqslant h_M^{(1)}(t) + \gamma^{-\frac{1}{2}}\left[\alpha\sigma(t) + \tfrac{1}{2}\beta\sigma^2(t)\right] + 3(\mu/\gamma)\,\sigma(t), \tag{3.9}$$

where $h_M^{(1)}$ is defined by

$$h_M^{(k)}(t) \equiv \max\left\{|h^{(k)}(s)| : 0 \leqslant s \leqslant t\right\}, \tag{3.10}$$

with t > 0 and k a non-negative integer. (Note that $|h(t)| \leqslant \sigma(t)$.)

Bounds for the spatial derivatives may also be deduced from (3.4). On differentiating (3.4) with respect to x we have that

$$\begin{aligned}\eta_{xt}(x,t) = {} & -\gamma^{-\frac{1}{2}}h'(t)\,e^{-x/\sqrt{\gamma}} + \int_0^\infty \tilde{K}_x(x,y).\,(\alpha\eta + \tfrac{1}{2}\beta\eta^2)(y,t)\,dy \\ & - \gamma^{-1}(\alpha\eta + \tfrac{1}{2}\beta\eta^2)(x,t) - \mu\gamma^{-\frac{3}{2}}h(t)\,e^{-x/\sqrt{\gamma}} - \frac{\mu}{\gamma}\eta_x(x,t) - \mu\int_0^\infty \tilde{H}_x(x,y)\,\eta(y,t)\,dy\,.\end{aligned} \tag{3.11}$$

Multiplying this equation by η_x(x,t) and using Hölder's inequality together with (3.6) it follows that

$$\tfrac{1}{2}\frac{\partial}{\partial t}\eta_x^2(x,t) + (\mu/\gamma)\,\eta_x^2(x,t) \leqslant M\,|\eta(x,t)|, \tag{3.12}$$

where $M \equiv \gamma^{-\frac{1}{2}}h_M^{(1)}(t) + 2\gamma^{-1}(\alpha\sigma(t) + \tfrac{1}{2}\beta\sigma^2(t)) + 2\mu\gamma^{-\frac{3}{2}}\sigma(t)$.

Gronwall's lemma implies that

$$|\eta_x(x,t)| \leq (M\gamma/\mu)(1 - e^{-\mu t/\gamma}) \equiv P_1(h_M^{(1)}(t), \sigma(t), t).$$

Thus, since $x \geq 0$ was arbitrary,

$$\|\eta_x(\cdot, t)\| \leq P_1(h_M^{(1)}(t), \sigma(t), t). \tag{3.13}$$

Note that $P_1(h_M^{(1)}(t), \sigma(t), t) \leq M \cdot \min\{t, \gamma/\mu\}$, so that P_1 is bounded by a polynomial linear in t and $h_M^{(1)}(t)$ and quadratic in $\sigma(t)$, having coefficients that are polynomials in α, β, μ and $\gamma^{-\frac{1}{2}}$. Moreover, the (explicit) dependence on t can be ignored for $t \geq \gamma/\mu$.

Bounds for higher-order spatial derivatives can be obtained inductively by similar kinds of arguments, leading to the following lemma.

Lemma 3.1. Let $T > 0$. Suppose that $h \in \mathscr{C}^1([0,T])$ and that $h(0) = 0$. Let η be the solution to (3.1) and let $h_M^{(1)}$ and σ be defined by (3.10) and(3.7) respectively. Then, for any positive integer k and for $t \in [0,T]$,

$$\|(\tfrac{\partial}{\partial x})^k \eta(\cdot, t)\| \leq P_k(h_M^{(1)}(t), \sigma(t), t),$$

where P_k can be bounded by a polynomial of degree k in $h_M^{(1)}(t)$ and $\min\{t, \gamma/\mu\}$, and of degree k+1 in $\sigma(t)$, having coefficients that are polynomials in α, β, μ and $\gamma^{-\frac{1}{2}}$

Comment. A polynomial $P(x_1, \ldots, x_n)$ is said to have degree ℓ_j in the variable x_j if P is a polynomial of degree at most ℓ_j in the variable x_j, when all the other variables are held fixed.

3.3. *Decay rates for the exact solution*

Lemma 3.2. Let $T > 0$ *and suppose that* $h \in \mathscr{C}^1([0,T])$, *with* $h(0) = 0$. *Let* η *be the solution to (3.1) and let* $h_M^{(0)}$, $h_M^{(1)}$ *and* σ *be defined by (3.10), (3.10) and (3.7) respectively. Then, for any real number* $r \in]0, \gamma^{-\frac{1}{2}}[$ *there is a function* $C = C(\sigma(t)) < \infty$ *such that*

$$|\eta(x,t)| \leq [(\mu/\gamma)\, h_M^{(0)}(t) + h_M^{(1)}(t)]\, C^{-1} \exp(Ct - rx)$$

for $t \in [0,T]$. *Here* $C(\xi) = (a/\gamma)\left[(\alpha + \frac{1}{2}\beta\xi) + \mu(1+\gamma r^2)/2\gamma^{1/2}\right]$, *where* $a = 2\gamma^{\frac{1}{2}}/(1-\gamma r^2)$.

Remarks. (i) This estimate says, in effect, that solutions to (3.1) represent waves which propagate with speed not exceeding C/r. When $\mu = 0$, the speed C/r is minimized when $r = (3\gamma)^{-\frac{1}{2}}$, so that $C/r \leq 3\sqrt{3}\ (\alpha + \frac{1}{2}\beta\sigma)$.

(ii) With $\mu = 0$, 'solitary wave' solutions to (3.1) of amplitude σ decay in space at a rate $r = \left[\gamma\,(1 + 3\alpha/\beta\sigma)\right]^{\frac{1}{2}}$ and propagate with speed $\alpha + \frac{1}{3}\beta\sigma$.

(iii) Similar results also apply when (3.1) is posed with non-zero initial data $\eta(x,0) = g(x)$, provided $g(0) = h(0)$, g' is square integrable and $|g(x)| \leq M e^{-r}$ for $x \geq 0$. The estimate is then modified by the addition of the term $M \exp(Ct - rx)$.

Proof. Let $X > 0$ and define a weighting function w such that

$$w(x) \equiv \begin{cases} e^{rx} & , \quad 0 \leq x \leq X \\ e^{rX} & , \quad x \geq X \end{cases} .$$

Set $v(x\ t) = \omega(x).\eta(x\ t)$ and multiply (3.4) by ω. It follows that

$$\begin{aligned} v_t(x,t) &= h'(t)e^{-x/\sqrt{\gamma}}\omega(x) + \int_0^\infty \widetilde{K}(x,y)\,\frac{\omega(x)}{\omega(y)}\,(\alpha +,\tfrac{1}{2}\beta\eta(y,t))v(y,t)\,dy \\ &\quad + \left(\frac{\mu}{\gamma}\right)\left[h(t)e^{-x/\sqrt{\gamma}}\omega(x) - v(x,t)\right] - \mu\int_0^\infty \widetilde{H}(x,y)\,\frac{\omega(x)}{\omega(y)}\,v(y,t)\,dy\ . \end{aligned} \tag{3.14}$$

When r is in the interval $]0,\gamma^{-\frac{1}{2}}[$, $e^{-x/\sqrt{\gamma}}.\omega(x) \leqslant 1$ for any $x \geqslant 0$. Therefore, after multiplying (3.14) by $v(x,t)$ and applying the Hölder inequality, we have that

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial t}v^2(x,t) + \frac{\mu}{\gamma}\,v^2(x,t) &\leqslant \Big\{|h'(t)| + \frac{\mu}{\gamma}|h(t)| + (\alpha+\tfrac{1}{2}\beta\sigma(t))\|v(\cdot,t)\|\int_0^\infty|\widetilde{K}(x,y)|\,\frac{\omega(x)}{\omega(y)}dy \\ &\quad + \mu\|v(\cdot,t)\|\int_0^\infty|\widetilde{H}(x,y)|\,\frac{\omega(x)}{\omega(y)}\,dy\Big\}\ |v(x,t)|\ . \end{aligned} \tag{3.15}$$

(Note that $\|v(\cdot,t)\| \leqslant \sigma(t)\,e^X < \infty$ for $t \in [0,T]$.)

From the definitions (3.5) it follows that

$$|\widetilde{K}(x,y)| \leqslant \gamma^{-1}\,e^{-|x-y|/\sqrt{\gamma}} \qquad \text{and} \qquad |\widetilde{H}(x,y)| \leqslant \gamma^{-\frac{3}{2}}\,e^{-|x-y|/\sqrt{\gamma}}\ .$$

Moreover, $w(x)/w(y) \leqslant \exp\left[(x-y)r\right]$ when $y \leqslant x$, and $w(x)/w(y) \leqslant 1$ when $y \geqslant x$, so that

$$\int_0^\infty e^{-|x-y|/\sqrt{\gamma}}\ \frac{\omega(x)}{\omega(y)}\,dy \ \leqslant\ (\gamma^{-\frac{1}{2}}-r)^{-1} + \gamma^{\frac{1}{2}} \ \equiv a_0\ . \tag{3.16}$$

Using these inequalities in (3.15) we see that

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial t}v^2(x,t) + \frac{\mu}{\gamma}\,v^2(x,t) &\leqslant \Big\{|h'(t)| + \frac{\mu}{\gamma}\,|h(t)| \\ &\quad + a_0\big[(\alpha+\tfrac{1}{2}\beta\sigma(t))/\gamma + \frac{\mu}{\gamma^{3/2}}\big]\,\|v(\cdot,t)\|\Big\}\ |v(x,t)|, \\ &\leqslant \Big\{\left(\frac{\mu}{\gamma}\right)h_M^{(0)}(t) + h_M^{(1)}(t) + \tilde{c}(t)\,\|v(\cdot,t)\|\Big\}\ |v(x,t)|\ , \end{aligned}$$

where $\tilde{c}(t) = a_0[(\alpha+\frac{1}{2}\beta\sigma(t))/\gamma + \mu/\gamma^{\frac{3}{2}}]$.

Write $S(t) \equiv \max\{\|v(\cdot,s)\| \ :\ 0 \leqslant s \leqslant t\}$. Then Gronwall's lemma implies, for any $t_0 < t \leqslant T$, that

$$\begin{aligned} |v(x,t)| &\leqslant \Big\{\left(\frac{\mu}{\gamma}\right)h_M^{(0)}(t) + h_M^{(1)}(t) + \tilde{c}(t)\,S(t)\Big\}\left(\frac{\gamma}{\mu}\right)\left(1 - e^{-\mu(t-t_0)/\gamma}\right) \\ &\quad + |v(x,t_0)|\,e^{-\mu(t-t_0)/\gamma}\ . \end{aligned}$$

But since x is an arbitrary point, it follows that

$$S(t) \leq \left\{\left(\frac{\mu}{\gamma}\right) h_M^{(0)}(t) + h_M^{(1)}(t) + \tilde{c}(t)\, S(t)\right\}\left(\frac{\gamma}{\mu}\right)\left[1 - e^{-\mu(t-t_0)/\gamma}\right] + S(t_0)\, e^{-\mu(t-t_0)/\gamma} .$$

Thus,

$$0 \leq \frac{S(t) - S(t_0)}{t - t_0} \leq \left\{\left(\frac{\mu}{\gamma}\right) h_M^{(0)}(t) + h_M^{(1)}(t) + \tilde{c}(t)\, S(t) - \left(\frac{\mu}{\gamma}\right) S(t_0)\right\} \times \left\{\frac{1 - \exp[-\mu(t-t_0)/\gamma]}{\mu(t-t_0)/\gamma}\right\}$$

and, on letting $t_0 \to t$, we have

$$S'(t) \leq \frac{\mu}{\gamma} h_M^{(0)}(t) + h_M^{(1)}(t) + \left[\tilde{c}(t) - \frac{\mu}{\gamma}\right] S(t) .$$

A further application of Gronwall's inequality gives

$$S(t) \leq \left[\frac{\mu}{\gamma} h_M^{(0)}(t) + h_M^{(1)}(t)\right] \frac{\exp[\tilde{c}(t) - (\mu/\gamma)t] - 1}{\tilde{c}(t) - (\mu/\gamma)} ,$$

so that

$$|\eta(x,t)| \leq \left[\frac{\mu}{\gamma} h_M^{(0)}(t) + h_M^{(1)}(t)\right] \frac{\exp[\tilde{c}(t) - (\mu/\gamma)t] - 1}{\tilde{c}(t) - (\mu/\gamma)} \cdot \frac{1}{w(x)} .$$

So far we have held X fixed, but if we now let $X \to \infty$ the conclusion of the lemma follows except that a is replaced by a_0 in the function $\tilde{c}(t)$. But, having deduced that $\eta(x,t)$ decreases exponentially in x we can repeat the above argument with $w(x) = \exp(rx)$ for all $x \geq 0$. We now know that $\|v(\cdot,t)\| = \|w(\cdot,t)\| < \infty$ and the argument using Gronwall's inequality is therefore valid. The improvement in the constant a comes because

$$\int_0^\infty e^{-|x-y|/\sqrt{\gamma}}\, e^{r(x-y)}\, dy \leq (\gamma^{-1/2} - r)^{-1} + (\gamma^{-1/2} + r)^{-1} \equiv a .$$

Using this estimate instead of (3.16) leads to the stated result.

4. THE NUMERICAL SCHEME

4.1 Spatial Discretization

The numerical scheme is based on the integral equation (3.4). The equation is first discretized in space, its right-hand side is evaluated and the resultant system of ordinary differential equations is integrated forward in time.

The spatial discretization is affected by approximating the integrals of (3.4) by the trapezoidal rule with derivative correction at the end points of the domain (see Davies & Rabinowitz 1967). Thus, truncating the half line $[0,\infty[$ and introducing a uniform partition of N+1 points, $\{0, \Delta x, 2\Delta x, \ldots, N\Delta x\}$, we have, for any sufficiently smooth function $V(x)$, the approximation

$$\begin{aligned}\int_{j\Delta x}^{k\Delta x} V(y)\,dy \approx I_{j,k}(V) \equiv \tfrac{1}{2}\Delta x\,(V(j\Delta x+) + V(k\Delta x-))\\ + \Delta x \sum_{i=j+1}^{k-1} V(i\Delta x) + \tfrac{1}{12}\Delta x^2\,(V'(j\Delta x+) - V'(j\Delta x-)),\end{aligned} \tag{4.1}$$

where j,k are natural numbers with $0 \leq j < k \leq N$. This approximation has fourth-order accuracy, provided V has four bounded continuous derivatives on the open interval $]j\Delta x, k\Delta x[$ (see §5.1 below).

In using (4.1) to approximate (3.5) we note that the function V(y) is of the form J(x,y).v(y), where v(y) is assumed to have four bounded, continuous derivatives on $]0,N\Delta x[$ and J(x,y) (which is used to symbolize either $\tilde{H}$ or $\tilde{K}$) has four bounded, continuous derivatives, as functions of y, on each of $]0,x[$ and $]x,N\Delta x[$. Thus, the approximation (4.1) is to be applied separately on each of these intervals and the sum is to be taken. When N is large enough it can be shown (see §5.1) that the contribution from the right-hand end point, $N\Delta x$, is negligibly small, so that terms arising there can be omitted from the numerical scheme. Therefore, if we denote

$J(i\Delta x, y)$ by $J_i(y)$, $0 \le i \le N$, it follows that

$$\int_0^{N\Delta x} J_i(y)\,v(y)\,dy \approx \tfrac{1}{2}\Delta x\left[J_i(0)v(0) + (J_i(i\Delta x-) + J_i(i\Delta x+))\,v(i\Delta x)\right] + \Delta x \sum_{j=1, j\neq i}^{N} J_i(j\Delta x)\,v(j\Delta x)$$

$$+ \tfrac{1}{12}\Delta x^2\left[(J_i(y)v(y))'\big|_{0+} - (J_i(y)v(y))'\big|_{i\Delta x-} + (J_i(y)v(y))'\big|_{i\Delta x+}\right. \tag{4.2}$$

$$\left(= I_{0,N}(J_i(y)v(y)) - \tfrac{1}{2}\Delta x\, J_i(N\Delta x)\,v(N\Delta x) + \tfrac{1}{12}\Delta x^2\,(J_i(y)v(y))'\big|_{N\Delta x-}\right).$$

If we further define $v_j \equiv v(j,x)$ and $J_{ij} \equiv J(i\,\Delta x,\ j\,\Delta y)$ and we introduce the particular forms for $\tilde{H}$ and $\tilde{K}$ we can (after some simplification) rewrite (4.2) in the form

$$\int_0^{N\Delta x} \tilde{H}_i(y)\,v(y)\,dy \approx e^{-i\Delta x/\sqrt{\gamma}}\left[\tfrac{1}{2}(\Delta x/\gamma^{3/2})\sum_{j=1}^{N} e^{-j\Delta x/\sqrt{\gamma}} v_j - \tfrac{1}{12}(\Delta x/\gamma)^2 v_0\right] + \Delta x \sum_{j=1}^{N} H_{ij} v_j + \tfrac{1}{12}(\Delta x/\gamma)^2\, v_i\ , \tag{4.3}$$

and

$$\int_0^{N\Delta x} \tilde{K}_i(y)\,v(y)\,dy \approx \tfrac{1}{2}(\Delta x/\gamma)\, e^{-i\Delta x/\sqrt{\gamma}}\left[v_0 + \sum_{j=1}^{N} e^{-j\Delta x/\sqrt{\gamma}} v_j + \tfrac{1}{6}\Delta x\, v'(0+)\right] + \Delta x \sum_{j=1}^{N} K_{ij} v_j - \tfrac{1}{12}(\Delta x^2/\gamma)\, v'(i\Delta x)\ , \tag{4.4}$$

with $K_{ii} \equiv 0$.

The continbous quantities $v'(0+)$ and $v'(i\,\Delta x)$ in (4.4) are still to be discretized. Since both these terms derive from the second-order correction terms to the trapezoidal ruleit is sufficient to approximate them only to second order to retain the overall fourth-order approximation of the integral. Thus, we write

$$v'(0+) \approx (-v_2 + 4v_1 - 3v_0)/2\Delta x\ , \quad v'(i\Delta x) \approx (v_{i+1} - v_{i-1})/2\Delta x\ ,$$

and, incorporating these approximations in (4.4), we have

$$\int_0^{N\Delta x} \tilde{K}_i(y)\,v(y)\,dy \approx \tfrac{1}{2}(\Delta x/\gamma)\, e^{-i\Delta x/\sqrt{\gamma}}\left\{\sum_{j=0}^{N} e^{-j\Delta x/\sqrt{\gamma}} v_j + \tfrac{1}{12}(-v_2 + 4v_1 - 3v_0)\right\} + \Delta x \sum_{j=1}^{N} K_{ij} v_j - \tfrac{1}{24}(\Delta x/\gamma)(v_{i+1} - v_{i-1})\,. \tag{4.5}$$

Using (4.3) and (4.5) we construct an approximation to (3.5) of the form

$$\eta_t \approx \underset{\sim}{F}(t, \alpha\eta + \tfrac{1}{2}\beta\eta^2) - \mu\,\underset{\sim}{G}(\eta)\,, \tag{4.6}$$

where $\underset{\sim}{F}$, $\underset{\sim}{G}$ are vector functions with, for example, the notation $\underset{\sim}{F} \equiv (F_1, \ldots, F_N)$. It is convenient for computational purposes to write each of $\underset{\sim}{F}$ and $\underset{\sim}{G}$ as the sum of two vectors, i.e.

$$\underset{\sim}{F}(t, \underset{\sim}{v}) \equiv \underset{\sim}{F}^1(t, \underset{\sim}{v}) + \underset{\sim}{F}^2(\underset{\sim}{v}) \quad , \quad \underset{\sim}{G}(\underset{\sim}{v}) \equiv \underset{\sim}{G}^1(\underset{\sim}{v}) + \underset{\sim}{G}^2(\underset{\sim}{v})$$

where

$$\left.\begin{aligned} F_i^1 &= e^{-i\Delta x/\sqrt{\gamma}}\left\{h'(t) + \tfrac{1}{2}(\Delta x/\gamma)\left[\sum_{j=0}^{N} e^{-j\Delta x/\sqrt{\gamma}} v_j + \tfrac{1}{12}(-v_2 + 4v_1 - 3v_0)\right]\right\} \\ &\qquad + \Delta x \sum_{j=1}^{N} K_{ij} v_j \quad , \quad \text{for } i = 1, 2, \ldots \quad , \\ F_i^2 &= \tfrac{1}{24}(\Delta x/\gamma)\begin{cases} -(v_{i+1} - v_{i-1}) \ , & \text{for } i = 1, 2, \ldots, N-1 \\ v_{N-1} \ , & \text{for } i = N \end{cases} \quad , \end{aligned}\right\} \tag{4.7}$$

$$\left.\begin{aligned} G_i^1 &= e^{-i\Delta x/\sqrt{\gamma}}\left\{-\left[\gamma^{-1} + \tfrac{1}{12}(\Delta x/\gamma)^2\right] v_0 + \tfrac{1}{2}(\Delta x/\gamma^{3/2}) \sum_{j=1}^{N} e^{-j\Delta x/\sqrt{\gamma}} v_j\right\} \\ &\qquad + \Delta x \sum_{j=1}^{N} H_{ij} v_j \quad , \quad \text{for } i = 1, 2, \ldots \quad , \\ G_i^2 &= \left[\gamma^{-1} + \tfrac{1}{12}(\Delta x/\gamma)^2\right] v_i \quad , \quad \text{for } i = 1, 2, \ldots, N. \end{aligned}\right\} \tag{4.8}$$

Here $\underset{\sim}{v} = (v_0, \ldots, v_N)$. Note that F_i^1, G_i^1 are defined $\forall\, i \geq 1$, even though they involve only $v_0, \ldots, v_N$.

4.2 An efficient computational procedure

Before discussing the discretization in time it is worthwhile to consider efficient ways of computing $\underset{\sim}{F}^1$ and $\underset{\sim}{G}^1$. Evaluating $\underset{\sim}{F}^1$ and $\underset{\sim}{G}^1$ directly would require $O(N^2)$ operations, although this can easily be reduced to O(N log N) operations, through the use of a fast convolution method. However, it is possible to view (4.6) as a difference approximation to (3.5) and reduce the computation of $\underset{\sim}{F}^1$ and $\underset{\sim}{G}^1$ to O(N) operations. To do this, we introduce a difference operator D^2 defined by

$$(D^2 \underset{\sim}{w})_i = w_i - (w_{i+1} - 2w_i + w_{i-1})/(e^{\Delta x/\sqrt{\gamma}} - 2 + e^{\Delta x/\sqrt{\gamma}}),$$

$$\equiv A w_i + B(w_{i+1} + w_{i-1}),$$

so that $A = 1 + 2B$.

The operator D^2 is effectively an infinite-order approximation to $(1-\gamma\partial_x^2)$, in the sense that when D^2 is applied to the integral kernel for the inverse of $(1-\gamma\partial_x^2)$, the Kronecker δ-function results, <u>exactly</u>. (To see this, let $w_i = e^{i\Delta x/\sqrt{\gamma}}$, $i \in \mathbb{Z}$. Then $D^2\underset{\sim}{w} \equiv 0$.) Thus, applying D^2 to $\underset{\sim}{F}^1$ and $\underset{\sim}{G}^1$ (and after some simplification) it follows, for $2 \leqslant i \leqslant N-1$, that

$$\left.\begin{aligned} [D^2 \underset{\sim}{F}^1(t,\underset{\sim}{v})]_i &= \tfrac{1}{2}(B\Delta x/\gamma)(v_{i+1} - v_{i-1}) \\ \text{and} \qquad [D^2 \underset{\sim}{G}^1(\underset{\sim}{v})]_i &= \tfrac{1}{2}(B\Delta x/\gamma^{3/2})\, e^{\Delta x/\sqrt{\gamma}}\, v_i . \end{aligned}\right\} \qquad (4.9)$$

In order to complete the system of equations we must calculate the values of $(D^2\underset{\sim}{F}^1)_i$, $(D^2\underset{\sim}{G}^1)_i$ for i = 1 and i = N. These yield, for i = 1,

$$\left.\begin{aligned} AF_1^1 + BF_2^1 &= -Bh'(t) + \tfrac{1}{24}(B\Delta x/\gamma)(13v_2 - 4v_1 - 9v_0) \\ \text{and} \qquad AG_1^1 + BG_2^1 &= (\gamma^{-1} + \tfrac{1}{12}(\Delta x/\gamma)^2)Bv_0 + (B\Delta x/\gamma^{3/2})\sinh(\Delta x/\sqrt{\gamma})\,v_1 , \end{aligned}\right\} (4.10)$$

and, for i = N,

$$\left.\begin{aligned} BF^{1}_{N-1} + AF^{1}_{N} &= -\tfrac{1}{2}(B\Delta x/\gamma)\,v_{N-1} + (-BF^{1}_{N+1})\,, \\ \text{and} \qquad BG^{1}_{N-1} + AG^{1}_{N} &= \tfrac{1}{2}(B\Delta x/\gamma^{3/2})\,e^{\Delta x/\sqrt{\gamma}}\,v_{N} + (-BG^{1}_{N+1})\,. \end{aligned}\right\} \quad (4.11)$$

We propose that $\underset{\sim}{F}^{1}$, $\underset{\sim}{G}^{1}$ be evaluated by solving the tridiagonal system of equations (4.9 - 4.11), which requires only O(N) operations. To solve these equations we must first evaluate the terms F^{1}_{N+1} and G^{1}_{N+1} that appear in (4.11), the calculations for which can be made explicitly using the formulae (4.7) , (4.8). (Note that such a computation involves only O(N) operations.) However, it can be shown (see §5.1) that the retention of the quantities F^{1}_{N+1} , G^{1}_{N+1} is of only exponentially small consequence and it is more convenient simply to discard them from the system (4.9 - 4.11). Let us denote the solution of the resulting set of equations (i.e. the ones neglecting F^{1}_{N+1} , G^{1}_{N+1} by $\tilde{\underset{\sim}{F}}^{1}$, $\tilde{\underset{\sim}{G}}^{1}$.

4.3. Temporal discretization

Let us denote the semi-discrete approximation to η by the vector function $\underset{\sim}{u}(t) = (u_0(t), u_1(t), \ldots, u_N(t))$, where $\underset{\sim}{u}$ is defined by

$$\left.\begin{aligned} u_0(t) &= h(t)\,, \quad t \geq 0\,, \\ \text{and} \qquad \dot{u}_i(t) &= \tilde{F}_i[t, (\alpha\underset{\sim}{u} + \tfrac{1}{2}\beta\underset{\sim}{u}^2)(t)] - \mu\tilde{G}_i[\underset{\sim}{u}(t)]\,, \quad i = 1, \ldots, N\,, \end{aligned}\right\} \quad (4.12)$$

and $\tilde{\underset{\sim}{F}} = \tilde{\underset{\sim}{F}}^{1} + \underset{\sim}{F}^{2}$, $\tilde{\underset{\sim}{G}} = \tilde{\underset{\sim}{G}}^{1} + \underset{\sim}{G}^{2}$. Here $\underset{\sim}{u}^{2}$ denotes the vector $(u_0^2, u_1^2, \ldots, u_N^2)$. Then if h is identified with u_0 wherever it appears in the definition of $\tilde{\underset{\sim}{F}}$ and $\tilde{\underset{\sim}{G}}$, the set of equations (4.12) may be written as a system of ordinary-differential equations

$$\dot{\underset{\sim}{u}} = \underset{\sim}{\mathcal{F}}(t, \underset{\sim}{u})\,, \qquad (4.13)$$

for the vector, $\underset{\sim}{u} = (u_1, u_2, \ldots, u_N)$. This set of equations can be shown (see §5.3) to have a solution on an interval $[0, T_0]$, where T_0 tends to infinity as both $\Delta x \to 0$ and $X_0 = N\Delta x \to \infty$.

The temporal discretization of (4.13) has been effected through a prediction-correction method which, in the present case, is efficient because the initial datum is zero and no start-up procedure is needed. The continuous quantity $h'(t)$ appearing in the definition of $\underset{\sim}{\tilde{F}}$ (cf. (4.7)) is calculated by the fourth-order central-difference formula

$$h'(n\Delta t) \approx dh^n = (h^{n-2} - 8h^{n-1} + 8h^{n+1} - h^{n+2})/(12\Delta t), \qquad (4.14)$$

where $h^n = h(n\Delta t)$, $n \in \mathbb{N}$. Let $\mathcal{F}^n(v)$ denote the function obtained by substituting dh^n for $h'(n\Delta t)$ in $\underset{\sim}{\mathcal{F}}$ at that point. Then we take the fully discrete approximation to η to be the vector function given by Moulton's method (cf. Isaacson & Keller 1966), namely

$$\left.\begin{aligned} \bar{u}^{n+1} &= u^n + \tfrac{1}{24}\Delta t\left[55\mathcal{F}^n(u^n) - 59\mathcal{F}^{n-1}(u^{n-1}) + 37\mathcal{F}^{n-2}(u^{n-2}) - 9\mathcal{F}^{n-3}(u^{n-3})\right] \\ &\text{and} \\ u^{n+1} &= u^n + \tfrac{1}{24}\Delta t\left[9\mathcal{F}^{n+1}(\bar{u}^{n+1}) + 19\mathcal{F}^n(u^n) - 5\mathcal{F}^{n-1}(u^{n-1}) + \mathcal{F}^{n-2}(u^{n-2})\right]. \end{aligned}\right\} (4.15)$$

Since the initial datum is presumed to be zero, we shall take u^0, u^{-1}, ... (and h^0, h^{-1}, ...) to be zero as the starting values for (4.15). It should be noted that there are no stability limitations on the size of Δt in (4.15) because Δx does not appear in the denominator of $\underset{\sim}{\mathcal{F}}$.

The error induced by using the above scheme to approximate the solution of (3.1) is $O(\Delta x^4 + \Delta t^4)$ and the same methods can be employed to develop schemes of arbitrary order of accuracy by using higher-order derivative corrections for the trapezoidal rule and higher-order prediction-correction methods. But before deducing the accuracy of the approximation we shall first describe some numerical tests made with the scheme.

4.4. Convergence Tests

The theoretical convergence rate of the scheme was checked by comparing numerical solutions for the propagation of a solitary wave with the 'exact' solution for the continuous equation. With $\mu = 0$ and for $x \in \mathbf{R}$, there is a family of exact solutions to (3.1a) of the form

$$\eta = \eta_0 \operatorname{sech}^2\left\{\left(\frac{\beta\eta_0}{12\gamma(\alpha+\frac{1}{3}\beta\eta_0)}\right)^{\frac{1}{2}}\left[x + x_0 - (\alpha+\tfrac{1}{3}\beta\eta_0)t\right]\right\} . \qquad (4.16)$$

where $\eta_0 > 0$ is the (maximum) amplitude of the wave and x_0 is a real constant. The wave propagates without change of form at a steady speed $(\alpha + \frac{1}{3}\beta\eta_0)$. The constant x_0 is a parameter used to 'offset' the solitary wave so that, at $t = 0$, the wave crest is located at $x = -x_0$; alternatively, the wave crest passes an observer stationed at $x = 0$ at time $t = x_0/(\alpha+\frac{1}{3}\beta\eta_0)$. Therefore, if at $x = 0$ we were to use (4.16) as the boundary data $h(t)$, we would have, in effect, an exact solution to (3.1). Of course this solution, as a function of t, is exact only on the whole real line but, because of the exponential decay in the 'tails' of (4.16), η can be made arbitrarily small at $t = 0$ (by choosing x_0 sufficiently large) so that for $t \in [0,\infty[$ the function (4.16) can provide a close approximation to an exact solution of (3.1). It should however be noted that such a truncation introduces an incompatibility at (0,0) between $\eta(0,t)$ and $\eta(x,0)$ and this may slightly pollute the numerical solutions.

Nevertheless, we have taken (4.16) as an 'exact' solution and have carried out a convergence test for the scheme, the results of which are given in table 4.1. Let $u_i(T)$ be the computed solution at time T and at $x = i\Delta x$, $0 \leq i \leq N$. Then the entries shown in table 4.1 are $\bar{E}_M(T) = \max\{|u_i(T) - \eta(i\Delta x,T)| : 0 \leq i \leq N\}$. The computations reported in table 4.1(a) were made with $\eta_0 = 0.25$,

Δ \ T	19.2	38.4	67.2	96.0	172.8	19.2	38.4	67.2
0.6	0.933(-4)	0.209(-1)	0.564(-1)	0.923(-1)	0.169			
ratio	12.4	14.3	13.1	12.5	10.0			
0.3	0.753(-5)	0.146(-2)	0.429(-2)	0.740(-2)	0.169(-1)			
ratio	13.6	15.6	15.5	15.7	16.3			
0.15	0.554(-6)	0.938(-4)	0.276(-3)	0.470(-3)	0.104(-2)	0.556(-7)	0.790(-4)	0.259(-3)
ratio	14.2	15.8	16.1	16.3	16.9	14.2	15.7	16.1
0.075	0.389(-7)	0.595(-5)	0.171(-4)	0.288(-4)	0.616(-4)	0.391(-8)	0.502(-5)	0.161(-4)
ratio						15.0	15.8	16.1
0.0375						0.260(-9)	0.317(-6)	0.100(-5)
$\|\eta\|_{L_\infty}$	0.011	0.250	0.250	0.250	0.250	0.11(-2)	0.250	0.250

(a) (b)

TABLE 4.1 The errors $\bar{E}_M(T)$ induced in integrating a solitary wave (4.16) with $\eta_0 = 0.25$, for a time T. $N\Delta x = 180.0$; $\Delta = \Delta t = \Delta x$; $\alpha = 1$, $\beta = 1.5$, $\mu = 0$, $\gamma = \frac{1}{6}$. (a) $\eta(0,0) = 0.1 \times 10^{-8}$, $x_0 \simeq 27.079$; (b) $\eta(0,0) = 0.1 \times 10^{-9}$, $x_0 \simeq 29.899$.

(which was roughly the largest wave amplitude encountered in the experiments) with x_0 chosen so that $\eta/\eta_0 = 0.1 \times 10^{-8}$ at (0,0), and with $\Delta t = \Delta x \ (\equiv \Delta)$. The choice of $\Delta t = \Delta x$ was made because preliminary tests suggested this was near the optimal choice, in terms of accuracy achieved for a given amount of work, and because it is sufficient to take $\Delta t/\Delta x$ = constant to check the convergence rate, if the error is proportional to $(\Delta t^4 + \Delta x^4)$. The domain used for these computations was approximately the same as that needed to make comparisons with the laboratory experiments.

It is seen in table 4.1(a) that, apart from the smallest time quoted, the errors decreased at approximately the 16:1 ratio expected of the scheme. At $t = 19.2$ the wave crest had not yet emerged from the 'wavemaker', so that the wave amplitudes were quite small (cf. the value of $\|\eta\|_{\ell_\infty}$ quoted in the table) and the influence of the truncation of the input waveform is reflected by convergence rates being rather smaller than expected for the scheme. With $\eta(0,0)/\eta_0$ chosen to be 0.1×10^{-9} the errors $\bar{E}_M$ (see table 4.1(b)) were of a similar form to those given in table 4.1(a). Indeed, it would appear that the differences derived from the different values of x_0 used in the two experiments: for the latter experiment the 'solitary' wave was centred approximately at $x = 45.70$ at $t = 67.20$, whereas for the former it was centred approximately at $x = 48.52$. However, when the error $\bar{E}_M$ was determined with the wave crest at 45.70, the errors for each experiment were nearly the same (for the cases $\Delta = 0.15, 0.075$).

A similar test of the convergence of the numerical scheme was made by comparing solutions with $\eta(X,t)$ for X fixed. If $u^j(X)$ is the computed solution at position X and at time $t = j\Delta t$, $0 \leq j \leq M$,

Δ	$E_1(X)$	$E_2(X)$	$E_M(X)$
0.6	0.286	0.178	0.165
ratio	15.6	13.6	13.8
0.3	0.183(-1)	0.131(-1)	0.120(-1)
ratio	16.6	15.1	15.5
0.15	0.110(-2)	0.870(-3)	0.774(-3)
ratio	16.9	15.8	16.0
0.075	0.651(-4)	0.551(-4)	0.485(-4)
$\Vert\eta\Vert$	1.09	0.426	0.250

TABLE 4.2 The errors E_1, E_2, E_M induced in integrating a solitary wave (4.16) with $\eta_0 = 0.25$, $X = 36.0$. $M\Delta t = 180.0$. $\Delta = \Delta t = \Delta x$; $\alpha = 1$, $\beta = 1.5$, $\mu = 0$, $\gamma = \frac{1}{6}$. $\eta(0,0) = 0.1\text{x}\ 10^{-8}$, $x_0 = 27.079$.

Δ	$E_1(15.0)$	$E_2(15.0)$	$E_M(15.0)$		$E_1(30.0)$	$E_2(30.0)$	$E_M(30.0)$
0.6	0.105	0.707(-1)	0.651(-1)		0.176	0.133	0.118
ratio	16.0	14.4	14.9		14.3	13.6	13.6
0.3	0.655(-2)	0.492(-2)	0.438(-2)		0.123(-1)	0.975(-2)	0.866(-2)
ratio	16.4	15.2	15.6		15.9	15.3	15.7
0.15	0.400(-3)	0.323(-3)	0.280(-3)		0.772(-3)	0.639(-3)	0.553(-3)
ratio	17.1	16.6	16.8		17.1	16.8	16.9
0.075	0.234(-4)	0.195(-4)	0.167(-4)		0.451(-4)	0.380(-4)	0.327(-4)
$\Vert\eta\Vert$	1.10	0.417	0.239		1.11	0.409	0.231

TABLE 4.3 A convergence table for the numerical scheme with $\mu \neq 0$. $M\Delta t = 75.0$. $\Delta = \Delta t = \Delta x$; $\alpha = 1$, $\beta = 1.5$, $\mu = 0.014$, $\gamma = \frac{1}{6}$.

then we have calculated

$$E_1(X) = \sum_{j=0}^{M} |u^j(X) - \eta(X, j\Delta t)|\,\Delta t \quad , \quad E_2(X) = \left\{ \sum_{j=0}^{M} [u^j(X) - \eta(X, j\Delta t)]^2 \Delta t \right\}^{1/2},$$

and

$$E_M(X) = max \left\{ |u^j(X) - \eta(X, j\Delta t)| : 0 \le j \le M \right\}.$$

The results of such a calculation for $\eta_0 = 0.25$ and $X = 36.0$ are given in table 4.2, and again a convergence rate of about 4 was obtained.

With $\mu \neq 0$, we do not know of an exact solution to the continuous equation, so the convergence rate of the scheme has had to be checked in a different way. To ascertain that the coding of the dissipative term was correct, experiments were run with the linear model (i.e. $\beta = 0$) with h(t) chosen to be sinusoidal in time and the decay rate of these waves was compared with that deduced from the dispersion relation. (The results of a test of this kind are described below in § 7.5)

Having checked that the dissipative term had been correctly coded, the convergence rate for the full equation (with $\beta = 1.5$) was estimated by taking the 'exact' solution to be the results from a computation made with a small value of Δ (i.e. $\Delta = 0.0375$) and comparing this solution with numerical solutions at larger values of Δ . Thus, using (4.16) at x = 0 for the boundary data h(t), with $\eta(0,0)/\eta_0$ chosen to be 0.1×10^{-8} (i.e. $x_0 \simeq 27.079$), and with $\mu = 0.014$ (the value used in the comparisons of § 7.3) the convergence rates, as shown in table 4.3, were again found to be about 4.

5. ERROR ESTIMATES FOR THE DISCRETE SCHEME

In this chapter we shall let c_i , $i = 1,2,\ldots$, denote real constants. Also, we shall assume that $\Delta t, \Delta x \leq 1$ so that the dependence of constants on *positive* powers of Δt and Δx can be ignored. The notation is the same as that used in §§ 3,4.

5.1. Spatial discretization errors

The error associated with the trapezoidal rule with derivative end correction, as given in (4.1), is:

Lemma 5.1. *If* V *has four bounded, continuous derivations on the open interval* $]j\Delta x, k\Delta x[$ *, then*

$$\left| \int_{j\Delta x}^{k\Delta x} V(y)\, dy - I_{j,k}(V) \right| \leq \frac{\Delta x^4}{384} \int_{j\Delta x}^{k\Delta x} |V^{(4)}(y)|\, dy \,.$$

This is a standard result (see, for example, Davis & Rabinowitz 1967).

The error arising from the use of the vector $\underset{\sim}{F}'$ can be estimated as the sum of a term proportional to Δx^4 and a term arising from our approximation at the right-hand extremity of the interval.

Lemma 5.2. *Suppose that* v *has four continuous derivatives on the interval* $[0, N\Delta x]$. *Let* $\underset{\sim}{v} = (v_0, \ldots, v_N)$, *where* $v_i \equiv v(i\,\Delta x)$. *Then*, *for* $i = 1,2,\ldots,N$,

$$\left| F_i(t,\underset{\sim}{v}) - h'(t) e^{-i\Delta x/\sqrt{\gamma}} - \int_0^{N\Delta x} \tilde{K}(i\Delta x, y)\, v(y)\, dy \right|$$

$$\leq c_1 \Delta x^4 \max\left\{ |v^{(j)}(x)| : x \in [0, N\Delta x],\ j = 0,\ldots,4 \right\} + c_2 \Delta x \max\left\{ |v_{N-k}| : k = 0,1,2 \right\},$$

where the constants c_1, c_2 *depend only on* γ .

Proof. By definition (see (4.1), (4.2), (4.7)) it follows, for $i = 1,\ldots,N-1$, that

$$F_i(t,\underset{\sim}{v}) = h'(t)\, e^{-i\Delta x/\sqrt{\gamma}} + I_{0,i}(\tilde{K}(i\Delta x,\cdot)\,v) + I_{i,N}(\tilde{K}(i\Delta x,\cdot)\,v)$$

$$- \frac{\Delta x^2}{12\gamma}\, e^{-i\Delta x/\sqrt{\gamma}}\left[v'(0) - \frac{(-v_2+4v_1-3v_0)}{2\Delta x}\right] + \frac{\Delta x^2}{12\gamma}\left[v'(i\Delta x) - \frac{(v_{i+1}-v_{i-1})}{2\Delta x}\right]$$

$$+ \tfrac{1}{2}\,\Delta x\, \tilde{K}(i\Delta x, N\Delta x)\, v_N + \tfrac{1}{12}\,\Delta x^2\,(\tilde{K}(i\Delta x,\cdot)\,v)'\Big|_{N\Delta x} .$$

The difference approximations in the fourth and fifth terms are less than

$$\tfrac{1}{3}\,\Delta x^2 \max\left\{|v^{(3)}(x)| : x \in [0, N\Delta x]\right\} .$$

The last term can be estimated as follows:

$$\begin{aligned}
|(\tilde{K}(i\Delta x,\cdot)\,v)'\big|_{N\Delta x}| &= |\tilde{K}'(i\Delta x, N\Delta x)\, v_N + \tilde{K}(i\Delta x, N\Delta x)\, v'(N\Delta x)| , \\
&\leq \gamma^{-3/2}|v_N| + \gamma^{-1}\,|v'(N\Delta x) - (v_{N-2} - 4v_{N-1} + 3v_N)/2\Delta x| \\
&\qquad + (\gamma^{-1}/2\Delta x)\,|v_{N-2} - 4v_{N-1} + 3v_N| , \\
&\leq (\Delta x^2/3\gamma)\max\{|v^{(3)}(x)| : x\in[0,N\Delta x]\} \\
&\qquad + (\gamma^{-3/2} + 2/(\gamma\Delta x))\max\{|v_{N-k}| : k = 0,1,2\} .
\end{aligned}$$

Combining these estimates, together with a direct estimate for the second last term, we have that

$$\begin{aligned}
&|F_i(t,\underset{\sim}{v}) - h'(t)\, e^{-i\Delta x/\sqrt{\gamma}} - I_{0,i}(\tilde{K}(i\Delta x,\cdot)\,v) - I_{i,N}(\tilde{K}(i\Delta x,\cdot)\,v)| \\
&\leq \frac{\Delta x^4}{12\gamma}\max\left\{|v^{(3)}(x)| : x\in[0,N\Delta x]\right\} + \Delta x\left(\frac{2}{3\gamma} + \frac{\Delta x}{12\gamma^{3/2}}\right)\max\left\{|v_{N-k}| : k = 0,1,2\right\} .
\end{aligned} \tag{5.1}$$

But lemma 5.1 implies that

$$E \equiv \left|I_{0,i}(\tilde{K}(i\Delta x,\cdot)\,v) + I_{i,N}(\tilde{K}(i\Delta x,\cdot)\,v) - \int_0^{N\Delta x} \tilde{K}(i\Delta x,y)\,v(y)\,dy\right|$$

$$\leq \frac{\Delta x^4}{384}\left[\int_0^{i\Delta x} |(\tilde{K}(i\Delta x,\cdot)\,v)^{(4)}(y)|\,dy + \int_{i\Delta x}^{N\Delta x} |(\tilde{K}(i\Delta x,\cdot)\,v)^{(4)}(y)|\,dy\right],$$

which can be estimated further through the use of Leibnitz's rule (together with (3.6)) and the Hölder inequality.

Thus, it follows that

$$E \leq c_3 \Delta x^4 \max\{|v^{(j)}(x)| : x \in [0, N\Delta x],\ j = 0, \ldots, 4\} ,$$

where c_3 depends only on γ . Then combining this estimate and (5.1) we have the required result for the case $i \neq N$. For the case $i = N$,

$$F_N(t, \underset{\sim}{v}) = h'(t) e^{-N\Delta x/\sqrt{\gamma}} + I_{0,N}(\tilde{K}(N\Delta x, \cdot) v) + \tfrac{1}{2}\Delta x\, \tilde{K}(N\Delta x, N\Delta x-) v_N$$
$$- \frac{\Delta x^2}{12\gamma} e^{-N\Delta x/\sqrt{\gamma}} \left[v'(0) - \frac{(-v_2 + 4v_1 - 3v_0)}{2\Delta x} \right] + \tfrac{1}{12}\Delta x^2 (\tilde{K}(N\Delta x, \cdot) v)'\Big|_{N\Delta x-} + \frac{\Delta x}{24\gamma} v_{N-1} .$$

The techniques used when $i \neq N$ apply in the same way in this case. (The constants here can be chosen to be the same as for the case $i \neq N$.)

A similar result can be established in relation to the vector $\underset{\sim}{G}$.

Lemma 5.3. *Suppose that* v *has four continuous derivatives on the interval* $[0, N\Delta x]$. *Let* $\underset{\sim}{v} = (v_0, \ldots, v_N)$, *where* $v_i \equiv v(i\Delta x)$. *Then, for* $i = 1, \ldots, N$,

$$| G_i(\underset{\sim}{v}) - \gamma^{-1} v_i + \gamma^{-1} e^{-i\Delta x/\sqrt{\gamma}} v_0 - \int_0^{N\Delta x} \tilde{H}(i\Delta x, y) v(y)\, dy |$$
$$\leq c_4 \Delta x^4 \max\{|v^{(j)}(x)| : x \in [0, N\Delta x],\ j = 0, \ldots, 4\} + c_5 \Delta x \max\{|v_{N-k}| : k = 0, 1, 2\} ,$$

where the constants c_4 , c_5 *depend only on* γ .

The proof of this lemma follows a similar pattern to that for lemma 5.2 and is therefore omitted.

The above lemmas can now be combined to give the following estimate.

Corollary 5.1. *Suppose that* v *has four continuous derivatives on the interval* $[0, N\Delta x]$. *Let* $\underset{\sim}{v} = (v_0, \ldots, v_N)$, *where* $v_i \equiv v(i\Delta x)$. *Then, for* $i = 1,2,\ldots,N$,

$$\left| \tilde{F}_i(t,\underset{\sim}{v}) - h'(t)e^{-i\Delta x/\sqrt{\gamma}} - \int_0^{N\Delta x} \tilde{K}(i\Delta x, y)\, v(y)\, dy \right|$$

$$+ \left| \tilde{G}_i(\underset{\sim}{v}) - \gamma^{-1} v_i + \gamma^{-1} e^{-i\Delta x/\sqrt{\gamma}} v_0 - \int_0^{N\Delta x} \tilde{H}(i\Delta x, y)\, v(y)\, dy \right|$$

$$\leq c_6 \Delta x^4 \max\left\{ |v^{(j)}(x)| : x \in [0, N\Delta x],\ j = 0, \ldots, 4 \right\}$$

$$+ c_7 \Delta x \max\left\{ |v_{N-k}| : k = 0, 1, 2 \right\}$$

$$+ e^{-(N+1)\Delta x/\sqrt{\gamma}} \left[|h'(t)| + c_8 \Delta x \sum_{j=0}^{N} e^{j\Delta x/\sqrt{\gamma}} |v_j| \right],$$

where $c_6 = c_1 + c_4$, $c_7 = c_2 + c_5$ *and* c_8 *is another constant depending only on* γ .

Proof. Define $\quad s_i = \sinh(i\Delta x/\sqrt{\gamma}) / \sinh[(N+1)\Delta x/\sqrt{\gamma}]$. $\hfill (5.2)$

Recall from §4.2 that $\underset{\sim}{\tilde{F}}^1$ and $\underset{\sim}{\tilde{G}}^1$ respectively differ from $\underset{\sim}{F}^1$ and $\underset{\sim}{G}^1$ only because the terms involving F^1_{N+1} and G^1_{N+1} were not retained. Thus, it follows from the definitions (4.10), (4.11) that

$$\tilde{F}_i(t,\underset{\sim}{v}) = F_i(t,\underset{\sim}{v}) - s_i F^1_{N+1}(t,\underset{\sim}{v}) \quad , \quad \tilde{G}_i(\underset{\sim}{v}) = G_i(\underset{\sim}{v}) - s_i G^1_{N+1}(\underset{\sim}{v}) \ , \tag{5.3}$$

for $i = 1,2,\ldots,N$. From the definition (4.7) of $\underset{\sim}{F}^1$ we see that

$$|F^1_{N+1}(t,\underset{\sim}{v})| \leq e^{-(N+1)\Delta x/\sqrt{\gamma}} \left\{ |h'(t)| + \tfrac{1}{2}(\Delta x/\gamma)\left[\sum_{j=0}^{N} e^{-j\Delta x/\sqrt{\gamma}} |v_j| + \tfrac{1}{4}|v_0| + \tfrac{1}{3}|v_1| + \tfrac{1}{12}|v_2| \right] \right\}$$

$$+ \Delta x \sum_{j=1}^{N} e^{(j-N-1)\Delta x/\sqrt{\gamma}} |v_j| \quad ,$$

$$\leq e^{-(N+1)\Delta x/\sqrt{\gamma}} \left\{ |h'(t)| + c' \Delta x \sum_{j=0}^{N} e^{j\Delta x/\sqrt{\gamma}} |v_j| \right\},$$

where $c' = 1 + \frac{2}{3}(\Delta x/\gamma)$. Similarly, it follows from (4.8) that

$$|G^1_{N+1}| \le c''\Delta x\, e^{-(N+1)\Delta x/\sqrt{\gamma}} \sum_{j=0}^{N} e^{j\Delta x/\sqrt{\gamma}} |v_j| ,$$

where c'' depends only on γ. Therefore, on defining $c_8 = c' + c''$, we have that

$$|\tilde{F}_i(t,\underset{\sim}{v}) - F_i(t,\underset{\sim}{v})| + |\tilde{G}_i(\underset{\sim}{v}) - G_i(\underset{\sim}{v})| \le e^{-(N+1)\Delta x/\sqrt{\gamma}}\left\{|h'(t) + c_8\Delta x \sum_{j=0}^{N} e^{j\Delta x/\sqrt{\gamma}}|v_j|\right\},$$

and the result follows from lemmas 5.2 and 5.3.

5.2. Lipschitz estimate for $\underset{\sim}{\mathcal{J}}$

In this section $\|\underset{\sim}{v}\|$ will be used to denote the ℓ_∞ norm of $\underset{\sim}{v}$; i.e. if $\underset{\sim}{v} = (v_1,\ldots,v_N)$, then $\|\underset{\sim}{v}\| = \max\{|v_i| : i = 1,\ldots,N\}$.

The map $\underset{\sim}{v} \mapsto \underset{\sim}{F}^1(t,\underset{\sim}{v})$ is an affine map, taking the form $\underset{\sim}{F}^1(t,\underset{\sim}{v}) = \underset{\sim}{L}(t) + \underset{\sim}{M}\,\underset{\sim}{v}$, say. The ℓ_∞ operator norm of $\underset{\sim}{M}$ (it is the maximum, absolute row sum of $\underset{\sim}{M}$) can be estimated as

$$\|\underset{\sim}{M}\| \le \max_{1\le i\le N+1}\left[e^{-i\Delta x/\sqrt{\gamma}}\left(\frac{\Delta x}{2\gamma}\right)\left(\sum_{j=0}^{N} e^{-j\Delta x/\sqrt{\gamma}} + \frac{2}{3}\right) + \frac{\Delta x}{2\gamma}\sum_{j=1}^{N} e^{-|i-j|\Delta x/\sqrt{\gamma}}\right],$$

$$\le \frac{3}{2}\left(\frac{\Delta x}{\gamma}\right)\sum_{j=0}^{N} e^{-j\Delta x/\sqrt{\gamma}} + \frac{1}{3}\left(\frac{\Delta x}{\gamma}\right) .$$

Since $\Delta x \sum_{j=1}^{N} e^{-j\Delta x/\sqrt{\gamma}} \le \sqrt{\gamma}$, it follows that $\|\underset{\sim}{M}\| \le 3/2\sqrt{\gamma} + 2/\gamma \equiv \frac{1}{2}c_{11}$. Therefore $\|\underset{\sim}{F}^1(t,\underset{\sim}{v}) - \underset{\sim}{F}^1(t,\underset{\sim}{w})\| = \|\underset{\sim}{M}(\underset{\sim}{v}-\underset{\sim}{w})\| \le \frac{1}{2}c_{11}\|\underset{\sim}{v}-\underset{\sim}{w}\|$. (Note that, here and below, the norm on $\underset{\sim}{v}-\underset{\sim}{w}$ is a norm on (N+1) - vectors whereas the remaining norms are taken on N-vectors.) Similarly, we have that

$$|F^1_{N+1}(t,\underset{\sim}{v}) - F^1_{N+1}(t,\underset{\sim}{w})| \le \frac{1}{2}c_{11}\|\underset{\sim}{v}-\underset{\sim}{w}\|$$

and a Lipschitz estimate on $\tilde{\underset{\sim}{F}}^1$ can be obtained thus:

$$\|\tilde{\underset{\sim}{F}}^1(t,\underset{\sim}{v}) - \tilde{\underset{\sim}{F}}^1(t,\underset{\sim}{w})\| \le \|\underset{\sim}{F}^1(t,\underset{\sim}{v}) - \underset{\sim}{F}^1(t,\underset{\sim}{w})\| + \|\tilde{\underset{\sim}{F}}^1(t,\underset{\sim}{v}) - F^1(t,\underset{\sim}{v}) - (\tilde{\underset{\sim}{F}}^1(t,\underset{\sim}{w}) - \underset{\sim}{F}^1(t,\underset{\sim}{w}))\|,$$

$$\le \|\underset{\sim}{F}^1(t,\underset{\sim}{v}) - \underset{\sim}{F}^1(t,\underset{\sim}{w})\| + |F_{N+1}(t,\underset{\sim}{v}) - F_{N+1}(t,\underset{\sim}{w})|,$$

$$\le c_{11}\|\underset{\sim}{v}-\underset{\sim}{w}\| .$$

The map $\underset{\sim}{v} \mapsto \underset{\sim}{F}^2(\underset{\sim}{v})$ is linear and its ℓ_∞ operator norm is bounded by $\Delta x/12\gamma$ (see (4.7)), so that an estimate for $\underset{\sim}{\tilde{F}} = \underset{\sim}{\tilde{F}}^1 + \underset{\sim}{F}^2$ is

$$\|\underset{\sim}{\tilde{F}}(t,\underset{\sim}{v}) - \underset{\sim}{\tilde{F}}(t,\underset{\sim}{w})\| \leq c_F \|\underset{\sim}{v} - \underset{\sim}{w}\| \text{ ,} \tag{5.4}$$

where $c_F = c_{11} + \frac{1}{12}\gamma$.

A similar argument can be applied to the map $\underset{\sim}{v} \mapsto \underset{\sim}{\tilde{G}}(\underset{\sim}{v})$ leading to an estimate of the form

$$\|\underset{\sim}{\tilde{G}}(\underset{\sim}{v}) - \underset{\sim}{\tilde{G}}(\underset{\sim}{w})\| \leq c_G \|\underset{\sim}{v} - \underset{\sim}{w}\| \text{ ,} \tag{5.5}$$

where c_G depends only on γ.

A combination of these two estimates can be used to obtain a Lipschitz estimate on $\underset{\sim}{\mathcal{J}}$ (see defined in (4.12) and (4.13)). Let $\underset{\sim}{\hat{v}}$ denote the vector $(v_1,\ldots,v_N)$, let $\underset{\sim}{v}$ denote $(h(t),v_1,\ldots,v_N)$ and let $\underset{\sim}{v}^2$ denote $(v_0^2,\ldots,v_N^2)$. Then, since $\underset{\sim}{\tilde{F}}(t,\underset{\sim}{v})$ is affine in $\underset{\sim}{v}$,

$$\begin{aligned}
\|\underset{\sim}{\mathcal{J}}(t,\underset{\sim}{\hat{v}}) - \underset{\sim}{\mathcal{J}}(t,\underset{\sim}{\hat{w}})\| &\leq \|\underset{\sim}{\tilde{F}}(t,\alpha\underset{\sim}{v} + \tfrac{1}{2}\beta\underset{\sim}{v}^2) - \underset{\sim}{\tilde{F}}(t,\alpha\underset{\sim}{w} + \tfrac{1}{2}\beta\underset{\sim}{w}^2)\| + \mu\|\underset{\sim}{G}(\underset{\sim}{v}) - \underset{\sim}{G}(\underset{\sim}{w})\| \text{ ,} \\
&\leq \alpha\|\underset{\sim}{\tilde{F}}(t,\underset{\sim}{v}) - \underset{\sim}{\tilde{F}}(t,\underset{\sim}{w})\| + \tfrac{1}{2}\beta\|\underset{\sim}{\tilde{F}}(t,\underset{\sim}{v}^2) - \underset{\sim}{\tilde{F}}(t,\underset{\sim}{w}^2)\| + \mu\|\underset{\sim}{G}(\underset{\sim}{v}) - \underset{\sim}{G}(\underset{\sim}{w})\| \text{,} \\
&\leq c_F(\alpha\|\underset{\sim}{v} - \underset{\sim}{w}\| + \tfrac{1}{2}\beta\|\underset{\sim}{v}^2 - \underset{\sim}{w}^2\|) + c_G\mu\|\underset{\sim}{v} - \underset{\sim}{w}\| \text{ ,} \\
&\leq \left[c_F(\alpha + \tfrac{1}{2}\beta\|\underset{\sim}{v} + \underset{\sim}{w}\|) + c_G\mu\right]\|\underset{\sim}{v} - \underset{\sim}{w}\| \text{ .}
\end{aligned}$$

Thus it follows that

$$\|\underset{\sim}{\mathcal{J}}(t,\hat{v}) - \underset{\sim}{\mathcal{J}}(t,\hat{\omega})\| \leq c_L(1 + \|\underset{\sim}{\hat{v}} + \underset{\sim}{\hat{w}}\|)\|\underset{\sim}{\hat{v}} - \underset{\sim}{\hat{w}}\| \text{ ,} \tag{5.6}$$

where c_L depends only on α,β,γ and μ. So we see that $\underset{\sim}{\mathcal{J}}$ is uniformly Lipschitz continuous in $\underset{\sim}{v}$ on bounded subsets of ℓ_∞.

5.3. Existence and bounds for the semi-discrete approximation

Let $\underset{\sim}{\eta}(t)$ represent the vector function $\eta_i(t) \equiv \eta(i\Delta x, t)$, $i = 1,\ldots,N$, where η is the solution to (3.1). Then, from corollary (5.1) and lemmas (3.1), (3.2) it follows that

$$\|\dot{\underset{\sim}{\eta}}(t) - \underset{\sim}{\mathcal{F}}(t,\underset{\sim}{\eta})\| \leq c_{12}\Delta x^4 P(h_M^{(1)}(t), \sigma(t), t) + e^{-(N+1)\Delta x/\sqrt{\gamma}} h_M^{(1)}$$
$$+ \frac{c_{13}\,\Delta x}{C}\left(\frac{\mu}{\gamma}\sigma(t) + h_M^{(1)}(t)\right)\left[e^{Ct-rN\Delta x} + e^{-(N+1)\Delta x/\sqrt{\gamma}}\sum_{j=0}^{N} e^{j\Delta x/\sqrt{\gamma}} e^{Ct-rj\Delta x}\right],$$

where $P(\xi,\sigma,t) = \max\{(1+P_i(\xi,\sigma,\tau))\,P_j(\xi,\sigma,\tau) : 0 \leq i < j \leq 4\}$, $C = C(\sigma(t))$, c_{12} and c_{13} depend only on α,β,γ and μ; and $0 < r < \gamma^{-1/2}$ (cf. lemma 3.2). This expression can be simplified by the use of the inequality

$$\Delta x\, e^{-(N+1)\Delta x/\sqrt{\gamma}}\sum_{j=0}^{N} e^{j\Delta x/\sqrt{\gamma} + Ct - rj\Delta x} \leq (\gamma^{-\frac{1}{2}} - r)^{-1} e^{Ct - r(N+1)\Delta x},$$

so that

$$\left.\begin{aligned}\|\dot{\underset{\sim}{\eta}}(t) - \underset{\sim}{\mathcal{F}}(t,\underset{\sim}{\eta})\| &\leq c_{12}\Delta x^4 P(h_M^{(1)}(t),\sigma(t),t) + h_M^{(1)}(t)e^{-(N+1)\Delta x/\sqrt{\gamma}}\\ &\quad + \frac{c_{14}}{C}\left[\frac{\mu}{\gamma}\sigma(t) + h_M^{(1)}(t)\right]e^{Ct-rN\Delta x},\\ &\equiv e_1(t) \quad [= e_1(h_M^{(1)}(t),\sigma(t),t,\Delta x, N\Delta x)],\end{aligned}\right\} \qquad (5.7)$$

and c_{14} depends only on α,β,γ,μ and r. Note that, by definition, e_1 is an increasing function of t.

Under the assumption that $h \in \mathcal{C}^1$, it follows from § 5.2 that $\underset{\sim}{\mathcal{F}}$ is locally Lipschitz continuous. Thus there is a unique solution $\underset{\sim}{u}(t)$ to (4.12) for $t \in [0, t_0]$ for some $t_0 > 0$. Suppose that T_0 is given by

$$T_0 = \sup\{t_0 \geq 0 : \underset{\sim}{u}(t) \text{ exists and } \|\underset{\sim}{u}(t) - \underset{\sim}{\eta}(t)\| \leq 1 \text{ for } t \in [0,t_0]\}. \qquad (5.8)$$

Since $\underset{\sim}{u}(0) = \underset{\sim}{\eta}(0) = \underset{\sim}{0}$, and both $\underset{\sim}{u}$ and $\underset{\sim}{\eta}$ are continuous, then $T_0 > 0$. We shall now obtain a lower bound for T_0 and show that $T_0 \to \infty$ as $\Delta x \to 0$ and $N\Delta x \to \infty$. For $t \in [0, T_0]$ it follows from (5.6) and (5.7) that

$$\|\dot{\underset{\sim}{u}}(t)-\dot{\underset{\sim}{\eta}}(t)\| \, (= \|\mathcal{F}(t,u(t))-\dot{\underset{\sim}{\eta}}(t)\|) \leq \|\underset{\sim}{\mathcal{F}}(t,\underset{\sim}{u}(t))-\underset{\sim}{\mathcal{F}}(t,\underset{\sim}{\eta}(t))\| + \|\underset{\sim}{\mathcal{F}}(t,\underset{\sim}{\eta}(t))-\dot{\underset{\sim}{\eta}}(t)\|,$$

$$\leq c_L(1+\|\underset{\sim}{u}(t)+\underset{\sim}{\eta}(t)\|)\,\|\underset{\sim}{u}(t)-\underset{\sim}{\eta}(t)\| + e_1(t),$$

$$\leq 2c_L(1+\sigma(t))\,\|\underset{\sim}{u}(t)-\underset{\sim}{\eta}(t)\| + e_1(t). \qquad (5.9)$$

Since $\frac{d}{dt}[max\{|u_i-\eta_i|\}] \leq max\{|\frac{d}{dt}(u_i-\eta_i)|\}$, except possibly on a set of zero measure, it follows from (a weak form of) Gronwall's lemma that

$$\|\underset{\sim}{u}(t)-\underset{\sim}{\eta}(t)\| \leq e_1(t)\left[e^{2c_L(1+\sigma(t))t}-1\right]\Big/\left[2c_L(1+\sigma(t))\right],$$
$$\equiv \psi(t), \qquad (5.10)$$

for $t \in [0,T_0]$.

However, if T_0 were such that $\psi(T_0) < 1$, it would contradict the maximality in the definition (5.8), as follows. In this case $\underset{\sim}{u}(t)$ is still defined for $t \in [T_0, T_0+t_1]$, $t_1 > 0$, because $\underset{\sim}{\mathcal{F}}$ is locally Lipschitz continuous; and $\|\underset{\sim}{u}(t)-\underset{\sim}{\eta}(t)\| \leq 1$, for $t \in [T_0, T_0+t_1]$, since $\underset{\sim}{u}$ and $\underset{\sim}{\eta}$ are continuous. Therefore $\psi(T_0) < 1$ cannot hold.

Since $e_1(t)$ and $\sigma(t)$ are non-decreasing in t, it follows that $\psi(t)$ is strictly increasing in t, as soon as $e_1(t) > 0$. Lemmas 3.1, 3.2 imply that σ is continuous and hence $\psi(t)$ is continuous. Also $\psi(t)\to\infty$ as $t\to\infty$. Thus it follows that $T_0 \geq \bar{T}$, where $\bar{T}$ is the unique solution of

$$\psi(\bar{T}) = 1. \qquad (5.11)$$

Note that, since $e_1(t)\to 0$ as $\Delta x\to 0$ and $N\Delta x\to\infty$ (with t fixed), $\bar{T}\to\infty$ as $\Delta x\to 0$ and $N\Delta x\to\infty$. Thus $\underset{\sim}{u}$ exists on an interval $[0,T_0]$

that becomes arbitrarily large as $\Delta x \to 0$ and $N\Delta x \to \infty$. Moreover

$$\|\underset{\sim}{u}(t)\| \leq \|\underset{\sim}{u}(t) - \eta(t)\| + \|\eta(t)\| \leq 1 + \sigma(t), \tag{5.12}$$

for $t \in [0, T_0]$. From now on we shall drop the distinction between T_0 and $\bar{T}$, and we will think of $\bar{T}$ as the upper limit of the time interval over which the above estimates are valid. Although $\bar{T} \leq T_0$ the advantage of using $\bar{T}$ is that it is determined by the equation $\psi(\bar{T}) = 1$, whereas T_0 is not.

The above estimates are valid under the assumption that $h \in \mathscr{C}^1([0,\bar{T}])$. We shall now derive bounds for the temporal derivatives of $\underset{\sim}{u}$ under the assumption that $h \in \mathscr{C}^k([0,\bar{T}])$, for some integer $k \geq 1$. These may be obtained directly from (4.13). Observe that $\underset{\sim}{\mathcal{F}}$ can be written in the form

$$\begin{aligned}\underset{\sim}{\mathcal{F}}(t,\underset{\sim}{v}) &= \underset{\sim}{\tilde{F}}\left[t, (\alpha h(t) + \tfrac{1}{2}\beta h^2(t), \alpha v_1 + \tfrac{1}{2}\beta v_1^2, \ldots, \alpha v_N + \tfrac{1}{2}\beta v_N^2)\right] - \mu \underset{\sim}{\tilde{G}}\left[h(t), v_1, \ldots, v_N\right] \\ &\equiv \underset{\sim}{\Gamma}(t) + \underset{\sim}{M}_F(\alpha \underset{\sim}{v} + \tfrac{1}{2}\beta \underset{\sim}{v}^2) - \mu \underset{\sim}{M}_G \underset{\sim}{v} ,\end{aligned} \tag{5.13}$$

where $\underset{\sim}{\Gamma}(t) = \underset{\sim}{\tilde{F}}\left[t, (\alpha h(t) + \tfrac{1}{2}\beta h^2(t), 0, \ldots, 0)\right] - \mu \underset{\sim}{G}\left[h(t), 0, \ldots, 0\right]$ and $\underset{\sim}{M}_F$, $\underset{\sim}{M}_G$ are matrices such that $\|M_F\| \leq c_F$ and $\|M_G\| \leq c_G$, as defined in (5.4), (5.5). (We recall the notation for the product $\underset{\sim}{u}\,\underset{\sim}{v}$, namely that $(uv)_i = u_i v_i$, $i = 1,2,\ldots,N$.) The vector $\underset{\sim}{\Gamma}(t)$ is given by (cf. definition 5.2)

$$\Gamma_i(t) = s_{N+1-i}\left[h'(t) + \frac{3\Delta x}{8\gamma}\left(\alpha + \tfrac{1}{2}\beta h(t)\right) - \mu\left(\gamma^{-1} + \tfrac{1}{12}\left(\frac{\Delta x}{\gamma}\right)^2\right)h(t)\right] + \begin{cases} \dfrac{\Delta x}{24\gamma}\left(\alpha h(t) + \tfrac{1}{2}\beta h^2(t)\right), & \text{if } i = 1 \\ 0 & \text{, if } i \geq 2 \end{cases} .$$

Thus $\left\|\left(\frac{d}{dt}\right)^k \underset{\sim}{\Gamma}(t)\right\| \leq q_k\left(h^{(0)}(t), \ldots, h^{(k)}(t)\right)$, where q_k is a quadratic polynomial with coefficients that are polynomials in

$\alpha, \beta, \gamma^{-1/2}, \mu$ and Δx , with rational coefficients.

From this partitioning of $\underset{\sim}{\mathcal{J}}$ we see that

$$\|\dot{\underset{\sim}{u}}(t)\| \leq q_1(h(t), h'(t)) + [\, c_F(\alpha + \tfrac{1}{2}\beta \|\underset{\sim}{u}(t)\| + \mu c_G\,]\, \|\underset{\sim}{u}(t)\| ,$$

$$\equiv Q_1(h(t), h'(t), \|\underset{\sim}{u}(t)\|),$$

where Q_1 is quadratic. Then, on differentiating (5.13) we have

$$\ddot{\underset{\sim}{u}} = \dot{\underset{\sim}{\Gamma}} + \underset{\sim}{M}_F(\alpha \dot{\underset{\sim}{u}} + \beta \underset{\sim}{u}\dot{\underset{\sim}{u}}) - \mu \underset{\sim}{M}_G \dot{\underset{\sim}{u}} ,$$

so that

$$\|\ddot{\underset{\sim}{u}}\| \leq q_2[h(t), h^{(1)}(t), h^{(2)}(t)] + [c_F(\alpha + \beta \|\underset{\sim}{u}(t)\| + c_G\mu]\, Q_1(h(t), h^{(1)}(t), \|\underset{\sim}{u}(t)\|),$$

$$\equiv Q_2(h(t), h^{(1)}(t), h^{(2)}(t), \|\underset{\sim}{u}(t)\|).$$

In this manner it can be shown inductively, together with the estimate (5.12) for $\|\underset{\sim}{u}(t)\|$, that

$$\|(\tfrac{d}{dt})^k \underset{\sim}{u}(t)\| \leq Q_k(h^{(0)}(t), \ldots, h^{(k)}(t), 1+\sigma(t)), \tag{5.14}$$

where Q_k is a polynomial of degree at most k+1,and $t \in [0, \bar{T}]$. Thus, it follows that $(\frac{d}{dt})^k \underset{\sim}{u}(0) = \underset{\sim}{0}$ if $h(0) = \ldots = h^{(k)}(0) = 0$, $k \geq 1$, and also that

$$\max\{\|(\tfrac{d}{dt})^k \underset{\sim}{u}(t)\| : t \in [0, \bar{T}]\} \leq Q_k(h_M^{(0)}(\bar{T}), \ldots, h_M^{(k)}(\bar{T}), 1+\sigma(\bar{T})).$$

Comment. Bounds for these temporal derivatives can also be obtained from (5.9) and (5.10). Proceding from that starting point, estimates can be obtained showing that $\|(\frac{d}{dt})^k(\underset{\sim}{u} - \underset{\sim}{\eta})(t)\| \to 0$ for any k,t, when $\Delta x \to 0$ and $N\Delta x \to \infty$.

5.4. Bounds for the fully discrete problem

Having shown in §5.3 that the semi-discrete approximation $\underset{\sim}{u}$ is close to η we shall now consider the fully discrete approximation as effected by the prediction-correction method (4.15). The following proposition is a direct adaptation of the results given in Isaacson & Keller (1966, see p. 388ff.).

Proposition. Let $\bar{T}, \Delta t > 0$ and let $\|\cdot\|$ be any norm on R^N. Suppose that $\underset{\sim}{y} = \underset{\sim}{y}(t) \in \mathscr{C}^5([-3\Delta t, \bar{T}], \mathbb{R}^N)$ is such that $\dot{\underset{\sim}{y}} = \underset{\sim}{f}(t, \underset{\sim}{y})$ on the interval $[-3\Delta t, \bar{T}]$, that $\underset{\sim}{y} \equiv \underset{\sim}{0}$ on $[-3\Delta t, 0]$ and that $\underset{\sim}{f}$ is Lipschitz continuous in $\underset{\sim}{y}$, with constant K, viz:

$$\|\underset{\sim}{f}(t,\underset{\sim}{u}) - \underset{\sim}{f}(t,\underset{\sim}{v})\| \leq K\|\underset{\sim}{u}-\underset{\sim}{v}\| \ , \ \forall\ \underset{\sim}{u}, \underset{\sim}{v} \in \mathbb{R}^N \text{ and for } t \in [-3\Delta t, \bar{T}] .$$

Let $\underset{\sim}{y}^n$, $n \geq 1$, be determined by

$$\left.\begin{aligned} \bar{\underset{\sim}{y}}^n &= \underset{\sim}{y}^{n-1} + \tfrac{1}{24}\Delta t\left(55\underset{\sim}{f}^{n-1} - 59\underset{\sim}{f}^{n-2} + 37\underset{\sim}{f}^{n-3} - 9\underset{\sim}{f}^{n-4}\right) + \Delta t\, \bar{\underset{\sim}{\theta}}^n , \\ \underset{\sim}{y}^n &= \underset{\sim}{y}^{n-1} + \tfrac{1}{24}\Delta t\left(9\bar{\underset{\sim}{f}}^n + 19\underset{\sim}{f}^{n-1} - 5\underset{\sim}{f}^{n-2} + \underset{\sim}{f}^{n-3}\right) + \Delta t\, \underset{\sim}{\theta}^n , \end{aligned}\right\} \tag{5.15}$$

where $\underset{\sim}{f}^j \equiv \underset{\sim}{f}(j\Delta t, \underset{\sim}{y}^j)$, $\bar{\underset{\sim}{f}}^j \equiv \underset{\sim}{f}(j\Delta t, \bar{\underset{\sim}{y}}^j)$ and $\underset{\sim}{y}^0 = \underset{\sim}{y}^{-1} = \underset{\sim}{y}^{-2} = \underset{\sim}{y}^{-3} = 0$. Suppose that the errors $\underset{\sim}{\theta}^n$ and $\bar{\underset{\sim}{\theta}}^n$ are such that

$$\|\underset{\sim}{\theta}^n\| + \tfrac{3}{8}K\Delta t\,\|\bar{\underset{\sim}{\theta}}^n\| \leq \theta \quad \text{for} \quad n \leq \bar{T}/\Delta t .$$

Then, for all $n \leq \bar{T}/\Delta t$, it follows that

$$\|\underset{\sim}{y}^n - \underset{\sim}{y}(n\Delta t)\| \leq \left[(b_3 + b_4 K\Delta t)\Delta t^4 \sup_{t\in[0,\bar{T}]} \|\underset{\sim}{y}^{(5)}(t)\| + \theta\right] \frac{(e^{c_d \bar{T}} - 1)}{c_d} ,$$

with $b_3 = 19/720$, $b_4 = 251/1920$ and $c_d = \frac{1}{12}K(17 + 30K\Delta t)$.

To apply this proposition to the scheme (4.15) we shall use the ℓ_∞ norm on $\mathbb{R}^N$ and let the errors $\bar{\underset{\sim}{\theta}}^n$, $\underset{\sim}{\theta}^n$ be

$$\left.\begin{aligned} \bar{\underset{\sim}{\theta}}^{n} &= \left[\frac{1}{24}\sum_{j=1}^{4}\bar{a}_j\, h'((n-j)\Delta t) - dh^{n-j}\right]\underset{\sim}{z} \\ \text{and}\qquad \underset{\sim}{\theta}^{n} &= \left[\frac{1}{24}\sum_{j=0}^{3} a_j\, h'((n-j)\Delta t) - dh^{n-j}\right]\underset{\sim}{z} \end{aligned}\right\} \qquad (5.16)$$

where $z_i = s_{N+1-i}$ (cf. (5.2)), dh^n is defined by (4.14) and $(\bar{a}_1 = 55\ ,\ \bar{a}_2 = -59\ ,\ \bar{a}_3 = 37,\ \bar{a}_4 = -9)$ and $(a_0 = 9,\ a_1 = 19,\ a_2 = -5,\ a_3 = 1)$.

Thus, $\underset{\sim}{\theta}^n$ and $\bar{\underset{\sim}{\theta}}^n$ can be estimated as

$$\|\underset{\sim}{\theta}^n\|\ ,\ \|\bar{\underset{\sim}{\theta}}^n\| \leqslant c_{15}\,\Delta t^4 \sup\left\{|h^{(5)}(t)| :\ t/\Delta t \in [n-6, n+2]\right\},$$

and c_{15} is simply a numerical constant.

A necessary condition for $\underset{\sim}{u}$ to be of class $\mathscr{C}^5$ is that $h \in \mathscr{C}^5$. Let us therefore assume that $h^{(0)}(0) = h^{(1)}(0) = \ldots = h^{(5)}(0) = 0$ and that $h \in \mathscr{C}^5([0, \bar{T}+2\Delta t])$ where $\bar{T}$ is the solution of (5.11), and define $h(t) = 0$, $\underset{\sim}{u}(t) = \underset{\sim}{0}$ for $t < 0$. Then $\underset{\sim}{u} \in \mathscr{C}^5([-\infty, \bar{T}])$ and $\dot{\underset{\sim}{u}}(t) = \underset{\sim}{\mathscr{F}}(t, \underset{\sim}{u}(t))$ for all $t \in [-\infty, \bar{T}]$. Moreover,

$$\max\left\{\|\underset{\sim}{\theta}^n\|, \|\bar{\underset{\sim}{\theta}}^n\| : n \leqslant \bar{T}\right\} \leqslant c_{15}\,\Delta t^4 \sup_{t\in[0,\bar{T}+2\Delta t]}\left\{|h^{(5)}(t)|\right\}.$$

Since the Lipschitz estimate on $\underset{\sim}{\mathscr{F}}$ is not a global estimate the above proposition cannot be used directly for the scheme (4.15). But an argument similar to the one used to prove the existence of $\underset{\sim}{u}$ in §5.3 can be employed to show that the proposition is applicable to $\underset{\sim}{\mathscr{F}}$ over a time interval $[0, T_1]$ where $T_1 \to \infty$ as $\Delta t \to 0$. However, because we are interested in deriving a posteriori error estimates for σ we shall follow a different argument.

Let $\tilde{T} \leqslant \bar{T}$ and set $\tilde{\sigma}(\tilde{T}) = \max\left\{\|\underset{\sim}{u}^n\|, \|\bar{\underset{\sim}{u}}^n\| : n \leqslant \tilde{T}/\Delta t\right\}$. Note that $\tilde{\sigma}$ depends implicity on Δt, Δx and N , but we shall view these as being fixed for the present. Regard $\tilde{\sigma}$ as a quantity computed by the above code. Therefore $\tilde{\sigma}$ is known, at least <u>a posteriori</u>. Define

$$B(\tau) = \left\{ \underset{\sim}{v} \in \mathbb{R}^N : \|\underset{\sim}{v}\| \leq \max\left\{\tilde{\sigma}(\tau),\ 1+\sigma(\tau)\right\}\right\}.$$

Thus, when $\tilde{T} < \bar{T}$, all the quantities $\underset{\sim}{u}^n$, $\underset{\sim}{\bar{u}}^n$ and $\underset{\sim}{u}(t)$ belong to $B(\tilde{T})$ for t, $n\,\Delta t \in [0,\tilde{T}]$. Then, define $\underset{\sim}{f}$ to be equal to $\underset{\sim}{\mathcal{F}}$ on $[0,\tilde{T}] \times B(\tilde{T})$ and such that $\underset{\sim}{f}(t,\underset{\sim}{v})$ is globally Lipschitz continuous in $\underset{\sim}{v}$ (for $t \in [0,\tilde{T}]$), with a Lipschitz constant not exceeding the Lipschitz constant for $\underset{\sim}{\mathcal{F}}$ restricted to $B(\tilde{T})$. This is possible because the temporal dependence and the $\underset{\sim}{v}$-dependence in $\underset{\sim}{\mathcal{F}}$ decouple (cf. 5.13). In particular, a bound for the Lipschitz constant for $\underset{\sim}{f}$ is afforded by

$$K(\tilde{T}) \equiv c_L\left(1 + 2\max\left\{\tilde{\sigma}(\tilde{T}),\ 1+\sigma(\tilde{T})\right\}\right).$$

Since $\underset{\sim}{u}^n$, $\underset{\sim}{\bar{u}}^n$ and $\underset{\sim}{u}(t)$, for t, $n\Delta t \in [0,\tilde{T}]$ may be viewed equivalently as having been generated either by $\underset{\sim}{\mathcal{F}}$ or $\underset{\sim}{f}$, the above proposition applies, yielding

$$\|\underset{\sim}{u}^n - \underset{\sim}{u}(n\Delta t)\| \leq c_{16}\left(1+\tfrac{3}{8}K(\tilde{T})\Delta t\right)\Delta t^4\left(\frac{e^{c_d\tilde{T}}-1}{c_d}\right)\left[\sup_{t\in[0,\tilde{T}]}\|\underset{\sim}{u}^{(5)}(t)\| + \sup_{t\in[0,\tilde{T}+2\Delta t]}|h^{(5)}(t)|\right],$$

for all $n \leq \tilde{T}/\Delta t$. Here, c_{16} is a numerical constant and

$$c_d = c_d\left[\sigma(\tilde{T}),\ \tilde{\sigma}(\tilde{T})\right] = \tfrac{1}{12}K(\tilde{T})\left[17 + 30K(\tilde{T})\Delta t\right].$$

Then, combining this estimate with (5.10) and (5.14) we have, for $0 \leq \tilde{T} \leq \bar{T}$ and for all $n \leq \bar{T}/\Delta t$ that

$$\begin{aligned}\|\underset{\sim}{u}^n - \underset{\sim}{\eta}(n\Delta t)\| \leq \psi(\tilde{T}) + c_{16}\left(1+\tfrac{3}{8}K(\tilde{T})\Delta t\right)\Delta t^4\left(\frac{e^{c_d\tilde{T}}-1}{c_d}\right)\Big[Q_5\big(h_M^{(0)}(\tilde{T}), \ldots \\ \ldots, h_M^{(6)}(\tilde{T}),\ 1+\sigma(\tilde{T})\big) + \sup_{t\in[0,\tilde{T}+2\Delta t]}|h^{(5)}(t)|\Big], \\ \equiv e_2 .\end{aligned} \tag{5.17}$$

The quantities $e_i(t) = e_i(h_M^{(0)}(t), \ldots, h_M^{(5)}(t), \sigma(t), \tilde{\sigma}(t), t, \Delta t, \Delta x, N\Delta x)$, $i \geq 2$, will be used to denote error expressions that tend to zero as $\Delta t, \Delta x \to 0$ and $N\Delta x \to \infty$. Thus, e_2 provides an estimate for the total error in the discrete scheme. In particular, for fixed $\tilde{T} > 0$, (5.17) shows that

$$|u_i^n - \eta(i\Delta x, n\Delta t)| \leq c_T(\Delta t^4 + \Delta x^4 + e^{-N\Delta x r}),$$

for all $n \leq \tilde{T}/\Delta t$, for $i = 1, \ldots, N$ and for any r such that $0 < r < \gamma^{-1/2}$. The constant c_T is independent of $\Delta t, \Delta x$ and N but depends on $\alpha, \beta, \gamma, \mu, r, h$ as well as $\tilde{T}$, and it is assumed that $\tilde{\sigma}(\tilde{T})$ stays bounded independently of $\Delta t, \Delta x$ and N.

However, the above estimates have the shortcoming that the quantity $\sigma(\tilde{T})$ appears exponentially on the right-hand side and that the a priori bound (3.3) for σ allows the possibility of growth in time. To obviate the possibility of such large growth rates, we shall derive an a posteriori bound on σ based on our knowledge of $\tilde{\sigma}$. Estimate (5.17) and the mean-value theorem imply that

$$\begin{aligned}
\max\{|\eta(x,t)| : x \in [0, N\Delta x], t \in [0, \tilde{T}]\} &\leq \max\{\|\underset{\sim}{\eta}(n\Delta t)\| : 0 \leq n \leq \tilde{T}/\Delta t\} \\
&\quad + \sqrt{2}(\Delta x + \Delta t)\max\{|\eta_x(x,t)| + |\eta_t(x,t)| : x \in [0, N\Delta x], t \in [0, \tilde{T}]\}, \\
&\leq \tilde{\sigma}(\tilde{T}) + e_2 + \sqrt{2}(\Delta x + \Delta t)\big[P_1(h_M^{(1)}(\tilde{T}), \sigma(\tilde{T}), \tilde{T}) + h_M^{(1)}(\tilde{T}) \\
&\quad + \gamma^{-\frac{1}{2}}(\alpha\sigma(\tilde{T}) + \tfrac{1}{2}\beta\sigma^2(\tilde{T})) + \tfrac{3\mu}{\gamma}\sigma(\tilde{T})\big], \\
&\equiv \tilde{\sigma}(\tilde{T}) + e_3 .
\end{aligned}$$

Then an upper bound for $|\eta(x,t)|$ for all $x \geq 0$ follows from lemma 3.2 and we have that

$$\begin{aligned}
\sigma(\tilde{T}) &\leq \tilde{\sigma}(\tilde{T}) + e_3 + \left(\tfrac{\mu}{\gamma} h_M^{(0)}(\tilde{T}) + h_M^{(1)}(\tilde{T})\right)\left(\frac{e^{C\tilde{T}} - 1}{C}\right) e^{-N\Delta x r} \\
&\equiv \tilde{\sigma}(\tilde{T}) + e_4 . \qquad (5.18)
\end{aligned}$$

As in the definition (5.11) of $\bar{T}$, there is a unique $T_2 > 0$ such that

$$e_4\left[h_M^{(0)}(T_2), \ldots, h_M^{(5)}(T_2), 1+\tilde{\sigma}(T_2), \tilde{\sigma}(T_2), T_2, \Delta t, \Delta x, N\Delta t\right] = 1 . \tag{5.19}$$

Furthermore, $T_2 \to \infty$ as $\Delta t, \Delta x \to 0$ and $N\Delta x \to \infty$, provided that $\tilde{\sigma}(t)$ remains finite for all finite $t \geqslant 0$. Thus, it follows from (5.18) that

$$\sigma(t) \leqslant 1 + \tilde{\sigma}(t) \quad \text{for} \quad t \in [0, T_2] \tag{5.20}$$

and $T_2 \leqslant \bar{T}$. Also we see that

$$|u_i^n - \eta(i\Delta x, n\Delta t)| \leqslant e_2\left[h_M^{(0)}(T), \ldots, h_M^{(5)}(T), 1+\tilde{\sigma}(T), \tilde{\sigma}(T), T, \Delta t, \Delta x, N\Delta x\right],$$

for $1 \leqslant i \leqslant N$, $n \leqslant T/\Delta t$ and $0 < T \leqslant T_2$, where e_2 is defined by (5.17) and T_2 is given by the solution to (5.19).

Thus, in summary, we have the following result.

Theorem. *Let* Δt *and* Δx *be positive parameters not exceeding one*. *Let* N *be a positive integer and let* $T > 0$. *Suppose that* *that* $h^{(i)}(0) = 0$ *for* $i = 0, 1, \ldots, 5$, *and that* η *is the solution* *to* (3.1). *Let* $\underset{\sim}{u}^n$ *be the solution of* (4.15) *and let* $\tilde{\sigma}(T) = max\{|u_i^n|: 1 \leqslant i \leqslant N, 1 \leqslant n \leqslant T/\Delta t\}$. *If* $T \leqslant T_2$, *as defined by (5.19), then*

$$max\left\{|\eta(i\Delta x, n\Delta t) - u_i^n| : 1 \leqslant i \leqslant N ,\ 1 \leqslant n \leqslant T/\Delta t\right\}$$

$$\leqslant c_{12}\, P(h_M^{(1)}(T),\ 1+\tilde{\sigma}(T),\ T)\, \Delta x^4$$

$$+ c_{14}\,(2+\tilde{\sigma}(T))\left[\frac{\mu}{\gamma}(1+\tilde{\sigma}(T)) + h_M^{(1)}(T)\right] \frac{e^{C(1+\tilde{\sigma}(T))T - rN\Delta x}}{C(1+\tilde{\sigma}(T))} \cdot \left(\frac{e^{2c_L(2+\tilde{\sigma}(T))T} - 1}{2c_L(2+\tilde{\sigma}(T))}\right)$$

$$+ c_{16}\left(1+\tfrac{3}{8}\tilde{K}(T)\Delta t\right)\left(\frac{e^{\tilde{c}_d(T)T} - 1}{\tilde{c}_d(T)}\right)\left[Q_5(h_M^{(0)}(T), \ldots, h_M^{(5)}(T), 2+\tilde{\sigma}(T)) + \sup_{t\in[0,T+2\Delta t]} |h^{(5)}(t)|\right]\Delta t^4,$$

where $\tilde{K}(T) = c_L(5+2\tilde{\sigma}(T))$ and $\tilde{c}_d(T) = \frac{1}{12}\tilde{K}(T)(17+30\tilde{K}(T)\,\Delta t)$. Here, $0 < r < \gamma^{-1/2}$; c_L, c_{12}, c_{14} and c_{16} are constants (introduced previously) which depend only on $\alpha, \beta, \gamma, \mu$ and r; P is defined in the proof of lemma 3.1 and in §5.3; C is defined in lemma 3.2; $h_M^{(i)}$, $i = 0,\dots,5$, are defined by (3.10); Q_5 is defined in §5.3 (cf. (5.14)).

Remarks. (i) The effects of round-off error can be incorporated into the above theorem as follows. Let the errors $\underset{\approx}{\theta}^n$, $\tilde{\underset{\approx}{\theta}}^n$ of (5.15) include the rounding error associated with the computation of f^n, y^n etc., at each time step. Suppose this additional error is bounded by θ_R (which will depend on N, Δx etc.). Then, using the proposition as stated, the final estimate in our theorem is modified simply by the addition of the term $\theta_R\left[exp(\tilde{c}_d(T)T)-1\right]/\tilde{c}_d(T)$.

(ii) A consequence of the *a posteriori* estimate is that we can replace the bound σ by $1+\tilde{\sigma}$ wherever it appears in the preceding estimates. However, σ and $\tilde{\sigma}$ may be small with respect to one, say $O(\varepsilon)$ and this replacement might not be a particularly good one. If we were to define T_0, T_1 by the unique solutions to $\psi(T_0)=\varepsilon$ and $e_4(T_2)=\varepsilon$ respectively, then $\sigma \leqslant \tilde{\sigma}+\varepsilon$ on $[0,T_2]$ so that σ may be replaced by $\tilde{\sigma}+\varepsilon$ wherever it occurs. In fact, we could define $\varepsilon = \max_t \tilde{\sigma}(t)$ and then σ can be replaced by 2ε on $[0,T_2]$. Note that, regardless of the size of $\varepsilon > 0$, T_0 and T_2 tend to infinity as $\Delta x, \Delta t \to 0$ and $N\Delta x \to \infty$.

6. EXPERIMENTAL APPARATUS AND PROCEDURE

6.1 Experimental apparatus

The experiments were carried out in a uniform channel of length 5.5 m and width 30 cm. One end of the channel was fitted with a plane beach of slope 1:10 ; at the other end there was a rigid plane flap which was used to generate the waves. The gap between the flap and the sides and bed of the channel was packed with foam plastic to restrict leakage past the wavemaker. In its rest position the flap was vertical and normal to the walls of the channel. It was supported by a horizontal shaft, the axis of which was normal to the walls of the channel at a height of about 1 m above the bed of the channel. The shaft was free to rotate about this axis. Since the water depth in the channel was only 3 cm for these experiments, the effective action of the paddle was very similar to that of a plane piston. The paddle was forced in an oscillatory motion by a long crank attached to an eccentric on the shaft of a synchronous motor. Thus, the frequency and amplitude of the paddle motion were fixed for any given experiment and the arrangement was such that the paddle could be set oscillating almost instantaneously under these conditions.

The walls and bed of the channel were made from plate glass. The width of the channel was uniform to within 0.01 cm and the bed was levelled so that it deviated from a mean horizontal plane by no more than 0.040 cm. (The r.m.s. variation in depth from the mean was 0.020 cm.) The levelling of the tank can be quite important, as any unevenness in the bed gives rise to reflected waves and systematic variations in depth lead to phase speeds different from those expected for a uniform channel. The walls of the channel were lined with an absorbent bandage to provide even wetting at the shoreline.

Wave heights were measured by means of proximity transducers placed near the surface of the water. (Briefly, the principle of the instrument is that these transducers form one plate of a capacitor, the liquid surface being the second plate. By determining the capacitance it is possible to infer the distance of the water surface from the transducer.) The output from these transducers was relayed to an ultraviolet chart recorder, giving a continuous record of the surface elevation. The frequency response of the system extended from d.c. to about 1 kHz. Since the sensitivity and range of a given transducer is related to its area,we have, by choosing the appropriate transducer, recorded wave amplitudes ranging between 0.005 cm and 0.5 cm with about the same relative accuracy over the entire range. The wave heights thus determined were accurate to within about 2% of the maximum recorded amplitude in any given run.

6.2. Experimental procedure

The tank was filled with water to roughly the desired depth and surface films were skimmed off. The water was then topped up until the level was within 0.001 cm of a reference level, set by the tip of a pointer gauge. For all the experiments to be described the mean water depth was 3.00 cm (the main uncertainty deriving from the unevenness in the bed, see above). A number of transducers (usually four) were then positioned along the channel, the distance of each transducer from the mean position of the wavemaker being known to within about 1 mm. Typically, the first transducer was placed about 15 to 20 cm from the wavemaker. On the basis of linear wavemaker theory (see Havelock 1929), we judged this distance to be well beyond the extent of the parasitic field of the wavemaker. The other transducers were then placed at distances of about 120 cm, 220 cm and 320 cm from the wavemaker.

When the surface of the water in the tank was free of disturbances the wavemaker was set in motion, executing sinusoidal oscillations at a fixed amplitude and frequency, and the water elevation at each of the transducers was recorded. The experiment was stopped when the wavefront reached the beach at the far end of the channel. All experiments to be described here were made at a fixed period of 0.6930 s (i.e. $\omega_0 = 0.5014$) for the motion of the paddle, but the amplitude of the motion was changed from experiment to experiment by adjusting the throw on the driving crank.

Under the above conditions the theoretical wavelength of infinitesimal waves is 36.00 cm, giving a wavelength-to-depth ratio of 12:1. (The reasons for this choice are outlined in §2.5.) It is instructive, then, to examine typical experimental conditions in relation to some of the theoretical assumptions for the model equations, as described in §2.1.

(a) The wave amplitude, ε , took values ranging between 0.002 and 0.2.

(b) The wavenumber, k_0 , was nominally 0.5234. The main reason for requiring that k be 'small' is that the dispersion relations for the model equations should be good approximations to the dispersion relation derived from the full linear theory (see equations (2.4)). For $k = 0.5234$ the phase speeds, ω/k, for the three models are

Model	Exact	KdV	M
ω/k	0.9580	0.9543	0.9562

so that the error in the phase speed for infinitesimal waves, arising from the use of model M, is less than 0.2%.

(c) The parameter S ($= \varepsilon(\lambda/d)^2$) took values between 0.4 and 36.

(d) The influence of surface tension is, to increase the phase speeds by about 0.1% (see Whitham 1974, p 403), which is smaller than the differences indicated in (b) above.

6.3 Comparison procedure

The analogue data representing the wave profiles were recorded at a chart speed of 300 mm s^{-1} so that, in one period of the wavemaker, roughly 200 mm of chart paper moved past the marking beam. A discretization of this signal was made by measuring the wave amplitudes at 4 mm intervals.† The peak-to-trough amplitude of the trace on the chart paper was adjusted to be about 60 to 70 mm (by suitably amplifying the output from the proximity gauge) and the displacement of the trace from its undisturbed position was measured to within about ± 0.3 mm. The above discretization corresponded to a temporal step of 0.2401 but preliminary tests suggested this would be too coarse for the degree of accuracy we would like for the numerical solutions. So, in order to use a time step of half this value, a (second-order) interpolation was made of the data obtained from the transducer nearest the wavemaker and the resulting data set was then used as the boundary datum,h,for the numerical computation. The initial datum g was taken to be zero for all experiments.

If the theoretical solution at the location X is given by $\eta(X,t)$, $t \in [0,T]$, and the observed wave amplitude at the same

† Digital recording and sampling facilities were not available and this restricted the range of experiments we were able to make in the present study.

position is denoted by $v(X,t)$, let us define an 'error' $E(\tau)$, $\tau \in \mathbb{R}$, between the two sets by

$$E(\tau) = \sum_{i=0}^{M} \{|\eta(X, i\Delta t - \tau) - v(X, i\Delta t)|\Delta t\} / \sum_{i=0}^{M} \{|v(X, i\Delta t)|\Delta t\}, \qquad (6.1)$$

where $M\Delta t = T$. The reason for introducing τ here is that $\inf\{E(\tau) : \tau \in \mathbb{R}\}$ gives essentially a measure of the difference in shape between the functions η and v, whereas the value of τ that realizes the infimum is effectively a phase error and can be used to provide a measure of the difference in the speed of propagation of the two waveforms. Thus, η and v could have a very similar 'shape' but give a relatively large value for $E(0)$ by virtue of only a small 'phase' error. So, in making comparisons between theoretical and experimental data, it is useful to evaluate $E(0)$, $\inf E$ and the 'phase' error.

The numerical solutions used for the comparisons to be described in §7 were obtained on a CYBER 175. With $\Delta t = 0.12005$ and $\Delta x = 0.15$, as used in the computations, the error $E(0)$ between an exact solitary-wave solution of the model equation and the computed solution, under conditions comparable to those of the experiment, was about 0.1%.

7. EXPERIMENTAL RESULTS

7.1 Damping coefficient

A determination of the damping along the channel was made from the steady wave field established after the wavemaker had been working for a long time. Although this situation greatly simplifies measurements of the wavefield, it adds the complication of having to identify the incident and reflected wave components. However, such a separation can be made without too much difficulty if there are no nonlinear effects present and if the waves are monochromatic. Indeed, for the same conditions as those used in the present experiments, Mahony & Pritchard (1980) measured a lapse rate $\eta^{-1}\eta_x$ of $0{\cdot}38\times10^{-2}$ at a wave amplitude of about 0.009. Prior to the present study a check of the decay rate was made at an amplitude of about 0.003 and this gave the same result as that found previously. Such a decay rate leads to a value for μ in (M*) of 0.014.

7.2 Two-dimensionality of the wavefield

The magnitude of the cross-channel variations of the wavefield were measured to see by how much the assumption of two-dimensionality of the wave motions was violated. This measurement was made by placing two transducers at different positions across the channel, but at the same distance along the channel from the wavemaker, and the difference between the signals from each of the transducers was formed.

The most important cross-channel structure was that of a transverse wave motion, an example of which is given in fig. 7.1. The waveform observed at the centre of the tank, at a distance 46.3d from the paddle, is shown in fig. 7.1(a) and the difference between the wave in the centre and that at a distance 5.9 cm from the side of the tank is shown in fig. 7.1(b). The transverse wave is seen to have an amplitude of about 4% of that of the longitudinal wave and a frequency twice that of the forcing frequency of the wavemaker. This is roughly the scale of the transverse motions at each of the observation points, at all amplitude settings. By moving one transducer across the tank relative to the other, it was also found to be representative of the size of the cross-channel variations. At the smaller wave amplitudes used in the experiments (ε less than about 0.01) a transverse motion was also evident, but the voltage differences

						I		II		III		IV	V
S	Station	x	$\sup\{\lvert\eta\rvert\}$	$\frac{-\log x}{\log \varepsilon}$	$\int_0^T \lvert\eta\rvert$	Expt Dissip. Model		Expt Inviscid Model		Expt Linear Model		Dissip (M*) Inviscid(M*)	(M*) Linear (M*)
0.38	B	70.9	0.0019	0.72	0.131	0.324 0.079	0.57	0.479 0.329	0.54	0.320 0.079	0.57	0.331	0.007
	C	103.8	0.0017	0.78	0.076	0.451 0.095	0.51	0.746 0.530	0.51	0.444 0.094	0.50	0.495	0.009
0.95	A	40.1	0.0053	0.74	0.336	0.100 0.098	0.06	0.287 0.286	0.04	0.101 0.097	0.06	0.185	0.017
	B	71.5	0.0045	0.85	0.208	0.103 0.092	0.06	0.429 0.426	0.07	0.099 0.092	0.06	0.338	0.017
	C	104.4	0.0043	0.93	0.108	0.191 0.106	0.22	0.549 0.509	0.22	0.183 0.110	0.20	0.480	0.022
4.5	A	40.1	0.0274	1.07	1.77	0.137 0.084	0.31	0.300 0.274	0.27	0.139 0.123	0.21	0.187	0.092
	B	71.5	0.0229	1.23	1.12	0.189 0.084	0.29	0.464 0.412	0.31	0.153 0.121	0.16	0.739	0.096
	C	104.5	0.0216	1.34	0.606	0.327 0.095	0.38	0.657 0.529	0.43	0.271 0.130	0.29	0.488	0.116
5.5	A	40.2	0.0335	1.13	1.98	0.109 0.080	0.26	0.262 0.247	0.19	0.132 0.131	0.06	0.186	0.111
	B	71.6	0.0285	1.31	1.27	0.113 0.081	0.14	0.379 0.360	0.15	0.115 0.114	-0.03	0.337	0.127
	C	104.4	0.0271	1.42	0.601	0.212 0.121	0.22	0.538 0.477	0.24	0.132 0.118	0.06	0.471	0.174
11.8	A	39.3	0.0687	1.47	2.15	0.144 0.097	0.32	0.274 0.230	0.34	0.223 0.210	-0.27	0.208	0.292
	B	70.9	0.0590	1.70	0.954	0.186 0.120	0.27	0.431 0.373	0.19	0.255 0.193	-0.34	0.357	0.398
	C	103.8	0.0527	1.85	1.70	0.420 0.107	0.51	0.845 0.546	0.66	0.200 0.189	-0.08	0.522	0.489
18.1	A	39.3	0.105	1.77	2.88	0.164 0.142	-0.23	0.279 0.272	-0.15	0.431 0.318	-1.01	0.236	0.368
	B	70.9	0.0917	2.06	1.40	0.162 0.162	-0.01	0.376 0.333	0.30	0.510 0.310	-0.80	0.393	0.456
	C	103.8	0.0747	2.24	1.62	0.205 0.188	-0.09	0.596 0.507	0.30	0.641 0.334	-0.66	0.560	0.596
26.3	A	13.4	0.209	1.53	10.8	0.352 0.346	-0.51	0.402 0.384	-0.73	0.760 0.645	-5.22	0.152	0.475
	B	26.7	0.149	1.93	10.8	0.197 0.193	-0.20	0.375 0.373	0.20	0.470 0.265	-1.94	0.245	0.403
	C	39.6	0.153	2.16	8.92	0.302 0.273	-0.47	0.452 0.449	-0.17	0.642 0.475	-1.70	0.284	0.551
35.9	A	39.3	0.201	2.64	4.91	0.235 0.221	0.21	0.401 0.318	0.63	0.851 0.523	-2.52	0.268	0.951
	B	70.9	0.155	3.07	2.82	0.331 0.128	0.53	0.880 0.358	1.18	1.17 0.447	-2.26	0.544	1.35
	C	103.8	0.124	3.34	2.92	0.512 0.136	0.63	1.51 0.580	1.46	1.37 0.540	-1.86	0.944	1.57
$\varepsilon(\lambda/d)^2$	Definition	x	$\sup\{\lvert\eta\rvert\}$	$\frac{-\log x}{\log \varepsilon}$	$\int_0^T \lvert\eta\rvert$	E(0) inf{E}	p(%)	E(0) inf{E}	p(%)	E(0) inf{E}	p(%)	E(0)	E(0)

TABLE 7.1 Detailed summary of comparisons made using the model

$$\eta_t + \alpha\eta_x + \beta\eta\eta_x - \mu\eta_{xx} - \gamma\eta_{xxt} = 0 \,, \qquad (M^*)$$

subject to $\eta(0,t)=h(t)$, $\eta(x,0)=0$, $x \geqslant 0$, $t \geqslant 0$. The entries for each column are defincd at the foot of the table. The errors, defined according to (6.1), are a comparison between the two sets of data indicated at the top of the appropriate column, with the abbreviations taking the following meanings.

Expt: Experimental data; Dissip. Model: $\alpha=1$, $\beta=\frac{3}{2}$, $\mu=0.014$, $\gamma=\frac{1}{6}$; Inviscid Model: $\alpha=1$, $\beta=\frac{3}{2}$, $\mu=0$, $\gamma=\frac{1}{6}$; Linear Model: $\alpha=1$, $\beta=0$, $\mu=0.014$, $\gamma=\frac{1}{6}$.

p represents the error in the 'phase' speed given as a percentage of the 'long-wave' speed, $\sqrt{gd}$; $p>0$ means the computed speed exceeds the experimental value.

between the two transducers were so small that they were only comparable with the noise level and it was therefore difficult to make any definitive statements about the frequency content of the transverse wave. However, the relative size of the transverse waves was certainly no greater than at the larger amplitudes.

The structure of the cross-channel motions was evidently rather complicated, being forced by the meniscus on the side walls or by the second harmonics of the longitudinal waves. We would expect the transverse motions to consist mainly of a mixture of wave modes of the form $\cos(m\pi y/b)$, where y is the cross-channel coordinate, b is the width of the channel and m is a natural number. Since waves at a frequency $2\omega_0$ satisfy the dispersion relation $\omega = (k\tanh k)^{\frac{1}{2}}$ at a value of $k \simeq 1.2043$, corresponding to a wavelength of 15.65 cm here, it would appear that the modes most easily excited should have been those with m = 3 or 4. Waves with m = 3 would have been able to radiate along the tank, whereas those with m = 4 would have been decaying modes.

7.3. The main comparisons

The main results of this study are summarized in table 7.1 and illustrated in figures 7.2 - 7.14. Several different kinds of tests have been carried out, as indicated in the table, but only a selection of the results are shown graphically. The eight experiments described in the table are defined by the parameter S, which ranged between 0.38 and 36. The 'stations' A,B,C are used to reference the locations of the transducer relative to the one used for the determination of h(t), the actual distances between the two transducers being given in the column headed 'x'. The wave amplitude ε is taken to be $\sup\{|h|\}$. The column headed $-\log x/\log\varepsilon$ indicates the position of the station expressed as a power of ε^{-1}. Henceforth we shall refer to the station at which the boundary data h(t) were measured as the 'boundary station'.

The comparisons given in columns I - III are the differences E, as defined in §6. The upper left-hand entry at each station is the difference E(0) and the entry below that is $\inf\{E(\tau) : \tau \in \mathbb{R}\}$ The entry to the right ~~of the dashed line~~ indicates the 'phase error' τ at which the infimum of E was realized, the error being expressed as a percentage of the time taken for a wave of speed 1.0 to reach the station. It is taken to be positive when the computed speed exceeded the experimental value. Column I shows the comparison between the experimental results and the wave amplitudes predicted by (M*). The second column shows the same kind of comparison, but no dissipative effects were included in the computations. For the results in the third column the nonlinear corrections were not included but dissipation was retained in the theoretical model.

The final two columns of the table show comparisons between different mathematical models, so only the difference E(0) is given. Thus, the second-last column shows the difference between the solutions with and without dissipative effects included and the last column gives an indication of the importance of the nonlinear term.

In the figures to be presented the unit used for the temporal axes is the time step Δt and that used for spatial coordinates is Δx. The diamond-shaped symbol represents the experimental data, in its discretized form,and the continuous curves are piecewise linear segments linking the computed values of the wave amplitudes at the mesh points.

The graph shown in figure 7.2 is the discretized form of the function

h used for the experiment at S = 5.5. The various comparisons for this experiment are shown in figures 7.3 to 7.5. Figure 7.3a shows the wave amplitudes as a function of x at four different times and figure 7.3b gives the comparisons between the nonlinear, dissipative version of (M*) and the experimental results. As given in the table, the relative differences E between the two functions were approximately 11%, 11% and 21% at stations A,B,C respectively but, after allowing for small phase-speed corrections of about 0.2% these differences were reduced to about 8%, 8% and 12% respectively. The importance of including the dissipative term is indicated by the results in figure 7.4, where the numerical solution is seen to differ markedly from the experimental results (cf. columns II and IV of table 7.1) On the other hand, the inclusion of the nonlinear term at this value of S is not so important, as shown in figure 7.5 (and see columns III,V of the table).

Note that, under these conditions, namely S = 5.5, the nonlinearity had the effect of modifying the waveform by about 17% at a 'distance' $\varepsilon^{-1.4}$ from the boundary station, whereas the dissipative effects had modified the waveform by 47% at the same distance along the channel. The nonlinear effects seem to have brought about only a slight flattening of the wave troughs and sharpening of the crests, a feature that can be seen by comparing figures 7.3b and 7.5.

An experiment for which the nonlinear effects were of only very minor importance is shown in figure 7.6. In this experiment S = 0.95 and the nonlinear term affected the waveform by only about 2%. The agreement between the theoretical prediction of (M*) and the observed waveform is not quite as good as for the results at S = 5.5, the main discrepancies apparently arising at the crests and troughs of the waves.

Similiar comparisons are shown in figure 7.7 (for S = 4.5), in figure 7.8 (S = 11.8) and in figure 7.9 (for the case S = 18.1). The experiment at S = 11.8 showed roughly the same kind of agreement as at the smaller values of S and this was confirmed by the quantitative comparisons. For the experiments at S = 4.5 and 5.5 the nonlinear term had had only a small beneficial effect on the theoretical prediction of the observations but, at S = 11.8, the inclusion of the nonlinear term provided a significantly better model than the linear dissipative theory (cf. columns I, III of the table). On the other hand, the inviscid model gave a very poor representation of all these experiments. Thus, while there was some advantage to be gained from retaining the nonlinear term under these conditions, it was far more important that the dissipative effects be taken into account.

The theoretical prediction of the experimental results at S = 18.1 was significantly worse than in the earlier cases. Whereas for all the previous experiments the difference E was less than about 10%, it was about 15% for the conditions at S = 18.1. One of the main reasons for the poorer agreement at S = 18.1 is that the theoretical speed of the leading wave appears to have been too large (see fig. 7.9), with the result that the phase correction needed to minimize $E(\tau)$ was quite different from that found for the earlier experiments. The contribution from the nonlinear terms at S = 18.1, which was quite large, is indicated in column V of the table.

Two experiments at yet larger amplitudes were made, one at S = 26.3 and the other at S = 35.9. For the experiment at S = 26.3 the stations A,B,C were located much nearer the boundary station than in the other experiments so that they would not lie beyond the (formal) range of validity of the model equation. The form of the boundary data h(t) for this experiment is given in fig. 7.10 and the structure of the wavefield along the channel at four times is shown in fig. 7.11a. The comparisons between the numerical solutions and the observed waveforms are shown in figs. 7.11b - 7.13. As indicated in the table, the agreement between the theoretical predictions and the experiment was not very close and the reason for this is apparent from the graphs. The experimental results indicate the presence of a substantial amount of second-harmonic component which is not nearly so strongly evident in the theoretical solutions of figure 7.11b (In retrospect, this property is also evident in the results shown in fig. 7.9 (S = 18.1), and fig. 7.8 (S = 11.8)). At station B the agreement is seemingly much better than at the other stations, but the reason for this appears to be that the phase of the second harmonic is such that it reinforces the trough and diminishes the crest of the observed waveform and so the agreement is probably fortuitous.

The experiment at S = 35.9 gave similar kinds of comparisons (see fig. 7.14) to those shown for the experiment at S = 26.3.

7.4 Assessment

The model appears to have given a fairly good description of the experiments at the smaller values of S, the differences being about 8 to 10%. To give more meaning to these comparisons it is worthwhile to examine some of the sources of error. There are two kinds of error involved: one arising from uncertainties in making the physical measurements and the other from not matching accurately the assumptions on which the model is based. For the present experiments, uncertainties in the physical measurements were not more than 2%, but since quantitative estimates of the other errors are not so easily made we shall attempt only a rough assessment of them. The non-uniformity of the waves across the channel (cf. § 7.2) was of the order of 4 or 5% of the wave amplitude. This feature could influence the results both through the inaccuracy of representing the initial data h(t) and through the error in making the comparisons at each of the stations A,B,C. In addition, there are uncertainties in the representation of the dissipative effects and deficiencies arising from the use of a one-dimensional model. Thus it does not seem as though we could expect closer agreement than the 8-10% found at the smaller values of S.

However, as S was increased, both the quantitative and the qualitative agreement between the experiments and the theory deteriorated, and it is of interest to ascertain why this should have been the case. There appeared to be three possible causes for the discrepancy.

(i) The dissipative effects were poorly modelled.

(ii) The presence of a non-negligible cross-wave component (cf. § 7.2).

(iii) The dispersion relation $\omega = (k \tanh k)^{1/2}$ was not very closely approximated by (M*) for wavenumbers near k_1. Thus, although the phase speeds of waves with wavenumbers near k_0 were closely approximated, the phase speeds of the shorter wavelengths evident in the experimental results were inaccurately represented by the model, and this feature could account for some of the disparities.

Without developing new theory or undertaking new experiments, it is not easy to account for (i) and (ii). We have, however, tried to make an appraisal of our modelling of the dissipation (see §7.5) and it is our view that this was not the main source for the discrepancies. It is, on the other hand, relatively straightforward to test the importance of (iii) (see § 7.6) and the tests suggest that this was the main source of weakness of the model with regard to the present experiments.

7.5 Modelling the dissipation

The comparisons described in § 7.3. indicate that the inclusion of dissipation is crucial if the model is to give a reasonable **description of the experimental results. therefore, in view of the** discussion of § 2.3, it seemed propitious to examine the sensitivity of the theory to different ways of modelling the dissipative effects.

In the comparisons of § 7.3 the theoretical solutions gave a reasonably good account of the experimental results at small values of S, but at larger values of S the agreement was not so good. A possible explanation of this is that wavenumbers different from k_0 were being dissipated at an incorrect rate and, in particular, the harmonics were likely to have been considerably overdamped because the dissipation was taken to be proportional to k^2. This would certainly be the case if all the damping occurred in the boundary layers (cf. eqn.(2.5)). In order to test the sensitivity of the model to the way the dissipative effects were represented we have examined the consequences of using some alternative models for the damping. For this purpose we shall work from the Ansatz that the entire damping at wavenumber k is proportional to $|k|^{1/2}$, as suggested by the boundary-layer theory, with the constant of proportionality, ρ_0, chosen to match the experimental decay rate at $k = k_0$. However it appears, at present, to be rather complicated to implement a numerical scheme to solve the initial- and boundary-value problem when the model equation includes the pseudo-differential operator whose symbol is $|k|^{1/2}$. Thus, for the purposes of this study, we have chosen to interpolate the function $\rho_0|k|^{1/2}$ by the polynomial $\nu+\mu k^2$. The interpolant used in §7.3, to be referred to as the $(0,k_0)$ interpolant, matched the magnitude of $\rho_0|k|^{1/2}$ at $k = 0$ and at $k = k_0$. But since this interpolant will dissipate waves with wavenumbers $k>k_0$ at a much faster rate than that implied by $\rho_0|k|^{1/2}$, we have also considered (k_0,k_1) interpolation of $\rho_0|k|^{1/2}$ where k_1 is the wavenumber corresponding to the frequency $2\omega_0$. This was done to provide a different representation of the damping of the wavemodes at the frequency $2\omega_0$ evident in the experimental results (e.g. see fig. 7.11b). (We have, incidentally, also examined the consequences of using Hermite interpolation of $\rho_0|k|^{1/2}$ by the function $\nu+\mu k^2$ at $k = k_0$; i.e. the magnitude and derivative of the functions were matched at k_0. But since the results were similar to those for the $(k_0\ ,\ k_1)$ interpolation we shall not describe them here.) Note that the terms $\nu+\mu k^2$ in the dispersion relation correspond to the terms $\nu\eta-\mu\eta_{xx}$ in the differential equation.

A series of numerical experiments were carried out using boundary data of the form $h(t) = \eta_0 \sin\omega_0 t$. To check that the dissipative terms had been correctly coded, a preliminary test was made, using the linear model $(\beta=0)$ for which the decay rate is known theoretically. To estimate the decay rate along the channel from the computed solutions, the amplitudes of the wave crests were found at a given time $(t = 172.8)$ and were plotted as a function of their distance from the boundary station. This graph (figure 7.15) shows that, except for a few crests near the front of the wavetrain, the amplitudes of the crests decreased at roughly the rate expected from the dispersion relation and that the two forms of dissipation gave similar results.

For comparison, we have also included in the graph the results of the same experiment with no dissipative effects (i.e. $\nu = 0 , \mu = 0$).

However, with $\beta = \frac{3}{2}$, the computed solutions differed significantly under the various representations of the dissipation, as illustrated in figure 7.16. This graph shows the computed wavefields, for $\eta_0 = 0.25$, at a time t = 172.8 corresponding roughly to the duration of a laboratory experiment. The comparison shown in this graph is that between the solutions obtained with the $(0 , k_0)$ interpolation of $\rho_0 |k|^{1/2}$ (dotted line) and with the (k_0 , k_1) interpolation of $\rho_0 |k|^{1/2}$ (full line). The substantial differences between these two solutions suggest that it could be very important to model the dissipative effects accurately.

To quantify the differences between these solutions we have evaluated the quantity $E = \sum_{j=0}^{N} \{ |\eta_1(j\Delta x,t) - \eta_2(j\Delta x,t)| \Delta x \} / \sum_{j=0}^{N} \{ |\eta_1(j\Delta x,t)| \Delta x \}$, for $x \in [0, N\Delta x]$, where η_1, η_2 are the functions being compared. Thus, for the comparison in fig. 7.16, the difference E = 0.313. A more complete list of comparisons, at various values of η_0 and at various times, is given in table 7.2. At small values of η_0 the differences were not too large, but with η_0 = 0.1 the differences had risen to about 10%.

Time η_0	57.6	115.2	172.8
0.005	0.034	0.039	0.041
0.050	0.048	0.068	0.078
0.100	0.065	0.098	0.119
0.250	0.136	0.248	0.313

TABLE 7.2 Values of E when $(0 , k_0)$ interpolation of $\rho_0 |k|^{1/2}$ was compared with (k_0 , k_1) interpolation of $\rho_0 |k|^{1/2}$, for $h(t) = \eta_0 \sin \omega_0 t$. The computations were made with $\Delta t = \Delta x = 0.15$.

The solutions given in fig. 7.16 suggest that the form of dissipation used for the comparisons of §7.3 probably dampened the larger wavenumbers too rapidly, which coul,d account for the theoretical solutions not yielding the shorter wavelength components apparent in the experimental results. Such a possibility was checked, in the case of the experiment at S = 26.3, by using the (k_0 , k_1) interpolation of $\rho_0 |k|^{1/2}$ to model the dissipative effects. A graph of the comparison is given in fig. 7.17, the error E(0) being 0.386, 0.283, 0.378 at the stations A,B,C respectively. These errors could be reduced to 0.368, 0.283, 0.363 with phase corrections of 0.84%, 0.07% and 0.90% at the respective stations. The agreement is slightly worse here than in §7.3. At stations A,C the amplitudes at the crests of the computed solutions were much smaller than those observed experimentally, whereas at station B the computed amplitudes of the crests were too large. But similar features to this were also evident in the solutions with no damping, i.e. $\nu = \mu = 0$ (see fig. 7.12), suggesting to us that the inaccurate model for the damping of the larger wavenumbers was probably not the main source of these discrepancies.

7.6 The approximation to the dispersion relation

First we examine the dispersion relations for the various models. These are shown graphically in fig. 7.18 where the shallow-water model ($\omega = k$) and (M) (as well as KdV) are compared with the 'exact' relation $\omega = (k \tanh k)^{1/2}$. By construction, these relations are all close at small values of k; at k = 0.5 it is evident that the shallow-water approximation is a poor model and at k = 1 all three models give a poor approximation to $\omega = (k \tanh k)^{1/2}$.

However, for the wavenumbers arising in our experiments, the equation

$$\eta_t + \alpha\eta_x + \beta\eta\eta_x + \nu\eta - \mu\eta_{xx} - \gamma\eta_{xxt} = 0 \qquad (\mathrm{M}\dagger)$$

can be used to provide a better interpolation of the 'exact' dispersion relation than that afforded by (M). This is achieved through a suitable choice of the parameters α, γ. Since the dominant wavenumbers appear to have been those corresponding to the frequencies ω_0 and $2\omega_0$, we have chosen α, γ so that the phase speeds for the linear form of (M†) (i.e. $\beta = 0$) coincided with those for the 'exact' theory at the wavenumbers k_0 and k_1. However, the displacement effects of the boundary layer lead, not only to a damping of the waves, but also to a correction in the phase speed of a wavemode (cf. eqn (2.5)). Therefore, taking the boundary-layer correction to the dispersion relation to be of the form suggested by the theory of Kakutani & Matsuuchi (1975), we have chosen to interpolate the dispersion relation

$$\omega = (k \tanh k)^{1/2} - \rho_0 (1 + i) |k|^{1/2}, \qquad (7.1)$$

with ρ_0 taken to be the empirical constant used for the comparisons in §7.3.

Under the conditions of the present experiments this interpolation gives $\alpha = 0.9898$, $\gamma = 0.1325$, for which values the real part of the dispersion relation for (M†) is shown in fig. 7.18(b), together with that for some of the other models. The theoretical solutions that result from the use of (M†) for the experiment at S = 26.3 are shown in fig. 7.19. The spatial form of the wavetrain, which is given in fig. 7.19(a), shows a number of qualitative differences from that obtained with (M*) (cf. fig. 7.11(a)), and the comparison between (M†) and the experimental results is given in fig 7.19(b). This comparison also shows a qualitative improvement in the prediction of

S	0.38	0.95	4.5	5.5	11.8	18.1	26.3	35.9
A		0.0 0.098 -0.03	0.117 0.091 0.24	0.082 0.076 0.12	0.133 0.077 0.35	0.075 0.075 -0.03	0.149 0.141	0.313 0.192 0.66
B	0.225 0.064 0.46	0.092 0.090 -0.05	0.143 0.088 0.20	0.063 0.059 0.05	0.157 0.104 0.26	0.091 0.048	0.156 0.156	0.514 0.161 1.01
C	0.326 0.061 0.41	0.114 0.103	0.244 0.100 0.28	0.103 0.069 0.11	0.367 0.090 0.47	0.107 0.077	0.194 0.177 -0.21	0.745 0.221 1.07

TABLE 7.3 The comparisons between the experimental results and (M+) with $\alpha = 0.9898$, $\beta = \frac{1}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$. The scheme is the same as for column I in table 7.1

the experimental results over the comparisons given in figs. 7.11(b) and (7.17). The quantitative comparisons for (M†), which gave differences for this experiment of 14%, 16% and 18% at the stations A,B,C respectively, are summarized in table 7.3. Indeed, (M†) represents all the experimental results to within about 8% except for the experiments at S = 26.3 and S = 35.9.

A graph of the comparison at S = 35.9 is given in fig. 7.20. The leading wave at each station is represented very well by the model (cf. the results of fig. 7.14 for (M*), where this was not the case), but the subsequent oscillations were modelled less accurately. Some other comparisons made with (M†) are given in appendix B.

Thus, it would appear that some of the major discrepancies between the predictions of the model and the experimental results originated in the poor theoretical representation of the phase speeds of some of the larger wavenumbers arising in the experiments.

8. RÉSUMÉ

The theoretical model predicted the experimental results, to as good an accuracy as could be expected, for the experiments made at values of S ranging up to 11.8. For these five experiments it was found that the inclusion of a dissipative term was much more important than the inclusion of the nonlinear term, although the inclusion of the nonlinear term was undoubtedly beneficial in describing the observations.

At larger values of S there were features of the experiments that were not predicted by the model. These features appear mainly to have been associated with harmonics (generated through nonlinear properties of the fundamental wavefield) having wavenumbers too large to be well represented by the small-wavenumber model. But by introducing a modification to the basic model that represented more accurately the phase speed of these harmonics, the description of the experimental results was significantly improved. On this basis, we feel that the original model would provide a good description of experiments in which the dominant wavenumber is much smaller than that used here, over a fairly wide range of values of the parameter S.

Finally, it is interesting to reflect briefly in a wider context on the implications of this study. We have used an efficient, unconditionally-stable, explicit scheme of fourth-order accuracy in both space and time to determine our numerical solutions. Computing these solutions to an accuracy of about 5% took roughly 10 s on the CYBER 175, which was about the same time as the physical run time of the laboratory experiment. Using a faster machine and with more efficient coding it is likely that the computation time could be reduced by a factor of about 10. Nevertheless, these numbers suggest that accurate computations with more complicated equations, such as those arising in weather forecasting, may be difficult to realize.

ACKNOWLEDGMENT

The first author is grateful for the hospitality of the Fluid Mechanics Research Institute, University of Essex, Essex, England.

APPENDIX A. Deficencies in an approximate procedure based on the pure initial-value problem

We wish to solve the pure initial-value problem

$$\eta_t + \eta_x + \tfrac{3}{2}\eta\eta_x - \tfrac{1}{6}\eta_{xxt} = 0 \quad , \quad x \in \mathbb{R} , \qquad \text{(M } \underline{\text{bis}}\text{)}$$

with the initial condition $\eta(x,0) = g(x)$. However, the initial datum g , to be determined empirically, is not easily obtained. Instead, a measurement of data $\eta(0,t) = \tilde{g}(t), t \geq 0$, is made and, to recover the intended problem, the function $\tilde{g}(t)$ is transformed to an 'equivalent' spatial representation $\tilde{g}(x)$ by the leading-order approximation $\eta_t + \eta_x = 0$ to (M). This transformation generates a small error, of order ε , in the representation $\tilde{g}(x)$ of the initial data $g(x)$, which ordinarily would not be important but, in the present example, is equivalent to the introduction of a forcing term on the right-hand side of (M) of comparable size to the nonlinear and the dispersive terms.

To illustrate the kinds of error that can arise in a practical case, let us consider the solitary-wave solution of (M), namely

$$\eta(x,t) = \eta_0 \operatorname{sech}^2\left\{\left(\frac{3\eta_0}{4+2\eta_0}\right)^{1/2}\left[x + x_0 - (1 + \tfrac{1}{2}\eta_0)t\right]\right\} , \qquad \text{(A1)}$$

where η_0 is the (maximum) wave amplitude and x_0 is a constant. Suppose that the measured data $\tilde{g}(t)$ is given by

$$\tilde{g}(t) = \eta_0 \operatorname{sech}^2\left\{\left(\frac{3\eta_0}{4+2\eta_0}\right)^{1/2}\left[x_0 - (1 + \tfrac{1}{2}\eta_0)t\right]\right\} ,$$

then, by choosing x_0 large enough, the solution to the initial- and boundary-value problem for (M) with

$$\eta(x,0) = 0 \quad , \quad \eta(0,t) = \tilde{g}(t)$$

(taking $\eta(0,0) = 0.1\times10^{-8}$), is a close approximation to (Al) for

$x, t > 0$ (cf. §3, Table 4.1).

If $\tilde{g}(t)$ is now transformed to an 'equivalent' spatial form, it follows that

$$\tilde{g}(x) = \eta_0 \operatorname{sech}^2 \left\{ \left(\frac{3\eta_0}{4+2\eta_0} \right)^{1/2} \left[x_0 + (1 + \tfrac{1}{2}\eta_0)\, x \right] \right\},$$

from which we see that $\tilde{g}(x)$ differs from the 'exact' form of g (the solution (A1) at time t = 0) in both its shape and phase, the phase difference between g(x) and $\tilde{g}(x)$ being $x_1 = \frac{1}{2}\eta_0 x_0 / (1 + \frac{1}{2}\eta_0)$. Notwithstanding the phase error, let us, for the time being, investigate the importance of the 'shape' error in $\tilde{g}$ by using $\tilde{g}(x-x_1)$ as the initial data for (M).

Thus, using a scheme similar to that described in §3 (which can also be analyzed in a similar way) we have solved numerically the pure initial-value problem (M) with $\eta(x,0) = \tilde{g}(x-x_1)$, which solution we denote by $\tilde{\eta}(x,t)$, and have compared $\tilde{\eta}$ with the 'exact' solution (A1).

The kinds of error that can arise in practice are shown in figure A1. Here the wave amplitude $\eta_0 = 0.25$ was chosen to correspond approximately to the largest amplitudes used in the laboratory experiments and the integration was carried on to about the same time as that occurring in the experiments. In figure A1 the dotted line represents the function $\tilde{\eta}$ and the full line represents the solitary-wave solution. Let $E = \sum_{j=0}^{N} \{ |\eta(j\Delta x, t) - \tilde{\eta}(j\Delta x, t)| \Delta x \} / \sum_{j=0}^{N} \{ |\eta(j\Delta x, t)| \Delta x \}$, where $x \in [0, N\Delta x]$, measure the difference between $\tilde{\eta}$ and the 'exact' solution. The initial difference between $\tilde{g}$ and g was approximately 0.111. At time t = 32.0 the approximate solution $\tilde{\eta}$ had developed a distinct oscillatory tail and the difference E had increased to 0.254. This difference then continued to increase with time, in a roughly linear fashion, taking values of 0.561 at t = 96.0 and 0.898 at t = 192.0. (The error E in integrating numerically a solitary wave of amplitude 0.25 under the conditions of this experiment was less than 0.55×10^{-3} at t = 192.0.) The figure shows how the tail developed by $\tilde{\eta}$ gradually separated from the leading wave. The speed of the leading crest of the oscillatory tail was approximately 0.9719 at t = 192.0. On the other hand, the leading wave of $\tilde{\eta}$ appeared to be evolving towards a solitary wave of the form (A1) with an amplitude of approximately 0.2139, as determined from a fourth-order interpolation of the discretized solution. For example, the speed of this wave differed from a solitary-wave solution (A1) of the same amplitude by less than 0.28×10^{-6} at t = 192.0 and the difference E between the two waveforms was less than 0.27×10^{-3}. (For this latter comparison the crest of the solitary-wave profile was chosen to coincide with that of the leading wave of $\tilde{\eta}$ and the domain ~~of integration~~ for the comparison was terminated at a distance a from the crest, where a was chosen so that the solitary waveform had decayed to 0.1×10^{-4} of its maximum amplitude).

Similar results were obtained with $\eta_0 = 0.1$, except that the initial error E at t = 0 was 0.048 and this degraded to 0.095 at t = 96.0 and 0.155 at t = 192.0. At t = 192.0 the amplitude of the leading wave of $\tilde{\eta}$ was 0.0971, but the wave was still undergoing significant modifications at this time.

It should be noted that the above comparisons underestimate considerably the actual errors arising with this method because we have removed the initial phase error induced by the approximate transformation of the data. (For example, with $\eta_0 = 0.1$ and

x_0 = 15.44 the error E between g(x) and $\tilde{g}(x)$ was 0.197, cf. the difference of only 0.048 between g(x) and $\tilde{g}(x-x_1)$.) With more general data it would not be so easy to eliminate this initial phase error.

APPENDIX B. More comparisons

We present here graphs of some of the comparisons listed in table 7.2 between the model (M†) and the experimental results. The graphs shown in figures C1-C5 are for the experiments at S = 0.95, 4.5, 5.5, 11.8 and 18.1 respectively. It can be seen from these figures that (M†) gives a much better overall representation of the results than that given by (M*), especially with regard to the leading waves of the train.

Then finally in fig. C6 we give, for comparison with figs 7.19b and 7.12 the results of a calculation made with no dissipation.

REFERENCES

BARNARD, B.J.S., MAHONY, J.J. & PRITCHARD, W.G. 1977 The excitation of surface waves near a cut-off frequency. *Phil. Trans. Roy. Soc. Lond. A.* **286**, 87.

BENJAMIN, T.B., BONA, J.L. & MAHONY, J.J. 1972 Model equations for long waves in nonlinear dispersive systems. *Phil. Trans. Roy. Soc. Lond.A.* **278**, 64.

BONA, J.L. & BRYANT, P.J. 1973 A mathematical model for long waves generated by wavemakers in nonlinear dispersive systems. *Proc. Camb. Phil.Soc.* **73**, 391.

BONA, J.L. & SMITH, R. 1975 The initial-value problem for the Korteweg-de Vries equation. *Phil. Trans. Roy. Soc. Lond. A.* **278**, 64.

DAVIS, P.J. & RABINOWITZ, P. 1967 *Numerical Integration.* Blaisdell, Waltham.

HAMMACK, J.L. 1973 A note on tsunamis: their generation and propagation in an ocean of uniform depth. *J. Fluid Mech.* **60**, 769.

HAMMACK, J.L. & SEGUR, H.1974 The Korteweg-de Vries equation and water waves. Part 2. Comparison with experiments. *J. Fluid Mech.* **65**, 289.

HAVELOCK, T.H. 1929 Forced surface waves on water. *Phil. Mag.*(F) **8**, 569.

ISAACSON, E.& KELLER, H.B. 1966 *Analysis of Numerical Methods*. John Wiley & Sons. New York.

KAKUTANI, T. & MATSUUCHI, K. 1975 Effect of viscosity on long gravity waves. *J. of the Phys. Soc. of Japan*. **39**, 237.

KEULEGAN, G.H. 1948 Gradual damping of solitary waves. *J. Res. Nat. Bur. Stand*. **40**, 487.

KORTEWEG, D.J. & DE VRIES, G. 1896 On the change of form of long waves advancing in a rectangular channel, and on a new type of long stationary waves. *Phil. Mag*. **39**, 422.

MAHONY, J.J. & PRITCHARD, W.G. 1980 Wave reflection from beaches. *J. Fluid Mech*. To appear.

MEI, C.C. & LIU, L.F. 1973 The damping of surface gravity waves in a bounded liquid. *J. Fluid Mech*. **59**, 239.

MEYER, R.E. 1972 Note on the longwave equations. *Univ. of Essex, Fluid Mech. Res. Inst. Rept. No. 23*.

MILES, J.W. 1967 Surface-wave damping in closed basins. *Proc. Roy. Soc. Lond. A.* **297**, 459.

PEREGRINE, D.H. 1966 Calculations of the development of an undular bore. *J. Fluid Mech.* **25**, 321.

WHITHAM, G.B. 1974 Linear and nonlinear waves. John Wiley & Sons. New York.

ZABUSKY, N.J. & GALVIN, C.J. 1971 Shallow-water waves, the Korteweg-de Vries equation and solitons. J. Fluid Mech. 47, 811.

FIGURE CAPTIONS

FIGURE 7.1. A tracing of the transducer voltage recorded at a distance 46.3 d from the paddle. The scaling for the ordinate has been made dimensionless; the frequency of the paddle was 0.6930 s ($\omega_0 = 0.5401$). (a) The wave profile at the centre of the channel. (b) The difference between a transducer at the centre and one placed at a distance 5.9 cm from the side of the channel.

FIGURE 7.2. The boundary data h(t) used for the calculation at S = 5.5.

FIGURE 7.3. The experiment at S = 5.5 is compared with (M*) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0.014$, $\gamma = \frac{1}{6}$. (a) Computed amplitudes as a function of x . (b) Temporal comparisons at stations A,B,C.

FIGURE 7.4. The experiment at S = 5.5 is compared with the inviscid version of (M*) ($\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0$, $\gamma = \frac{1}{6}$).

FIGURE 7.5. The experiment at S = 5.5 is compared with the linear version of (M*) ($\alpha = 1$, $\beta = 0$, $\mu = 0.014$, $\gamma = \frac{1}{6}$).

FIGURE 7.6. The experiment at S = 0.95 compared with (M*) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0.014$, $\gamma = \frac{1}{6}$.

FIGURE 7.7. The experiment at S = 4.5 is compared with (M*) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0.014$, $\gamma = \frac{1}{6}$.

FIGURE 7.8. The experiment at S = 11.8 is compared with (M*) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0.014$, $\gamma = \frac{1}{6}$.

FIGURE 7.9. The experiment at S = 18.1 is compared with (M*) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0.014$, $\gamma = \frac{1}{6}$.

FIGURE 7.10. The boundary data h(t) used for the calculation at S = 26.3.

FIGURE 7.11. The experiment at S = 26.3 is compared with (M*) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0.014$, $\gamma = \frac{1}{6}$. (a) Computed amplitudes as a function of x . (b) Temporal comparisons.

FIGURE 7.12. The experiment at S = 26.3 is compared with the inviscid version of (M*) ($\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0$, $\gamma = \frac{1}{6}$).

FIGURE 7.13. The experiment at S = 26.3 is compared with the linear version of (M*) ($\alpha = 1$, $\beta = 0$, $\mu = 0.014$, $\gamma = \frac{1}{6}$).

FIGURE 7.14. The experiment at S = 35.9 is compared with (M*) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\mu = 0.014$, $\gamma = \frac{1}{6}$.

FIGURE 7.15. Computed amplitudes of wave crests as a function of distance from the boundary station for linear models ($\beta = 0$), with boundary data $h(t) = \eta_0 \sin \omega_0 t$. Time t = 172.8. O: inviscid model ($\nu = 0$, $\mu = 0$); ●: $\nu = 0$, $\mu = 0.01$; △: $\nu = 0.24 \times 10^{-2}$, $\mu = 0.11 \times 10^{-2}$. The slopes of the straight lines are $-(\nu + \mu k_0^2)$. The computations were made with $\Delta t = \Delta x = 0.15$.

FIGURE 7.16. Computed wavefields at time $t = 172.8$ for nonlinear models ($\beta = \frac{3}{2}$) with boundary data $h(t) = 0.25 \sin \omega_0 t$ . ... : $\nu = 0$, $\mu = 0.014$ ((0, k_0) interpolation of $\rho_0 |k|^{1/2}$); ——— : $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$ ((k_0, k_1) interpolation). The computations were made with $\Delta t = \Delta x = 0.15$.

FIGURE 7.17. The experiment at $S = 26.3$ is compared with (M+) (see § 7.6) when $\alpha = 1$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = \frac{1}{6}$.

FIGURE 7.18. Graphs of the linear dispersion relations for various models. (a) — — —: shallow-water model, $\omega = k$; ———:'exact' relation, $\omega = (k \tanh k)^{1/2}$; --------- : model (M), $\omega = k/(1+\frac{1}{6}k)$; —·—·—: (KdV), $\omega = k(1-\frac{1}{6}k)$. (b) Magnified version of (a). — — — : $\omega = k$; ——— : $\omega = (k \tanh k)^{1/2}$;—·—: $\omega = 0.9898k/(1+0.1325k)$; .... : $\omega = k/(1+\frac{1}{6}k)$.

FIGURE 7.19. The experiment at $S = 26.3$ is compared with (M+) when $\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$. (a)Computed amplitudes as a function of x. (b) Temporal comparisons.

FIGURE 7.20. The experiment at $S = 35.9$ is compared with (M+) when $\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$.

FIGURE A1. The solitary-wave solution of (M) (full line) is compared with the solution to (M) with the initial data $\eta(x,0) = \tilde{g}(x-x_1)$ (dotted lines). The amplitude $\eta_0 = 0.25$; the computations were made with $\Delta t = \Delta x = 0.16$.

FIGURE B1. The experiment at $S = 0.95$ is compared with (M+) when $\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$.

FIGURE B2. The experiment at $S = 4.5$ is compared with (M+) when $\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$

FIGURE B3. The experiment at S = 5.5 is compared with (M+) when $\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$.

FIGURE B4. The experiment at S = 11.8 is compared with (M+) when $\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$.

FIGURE B5. The experiment at S = 18.1 is compared with (M+) when $\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0.340 \times 10^{-2}$, $\mu = 0.168 \times 10^{-2}$, $\gamma = 0.1325$.

FIGURE B6. The experiment at S = 26.3 is compared with the inviscid version of (M+) ($\alpha = 0.9898$, $\beta = \frac{3}{2}$, $\nu = 0$, $\mu = 0$, $\gamma = 0.1325$).

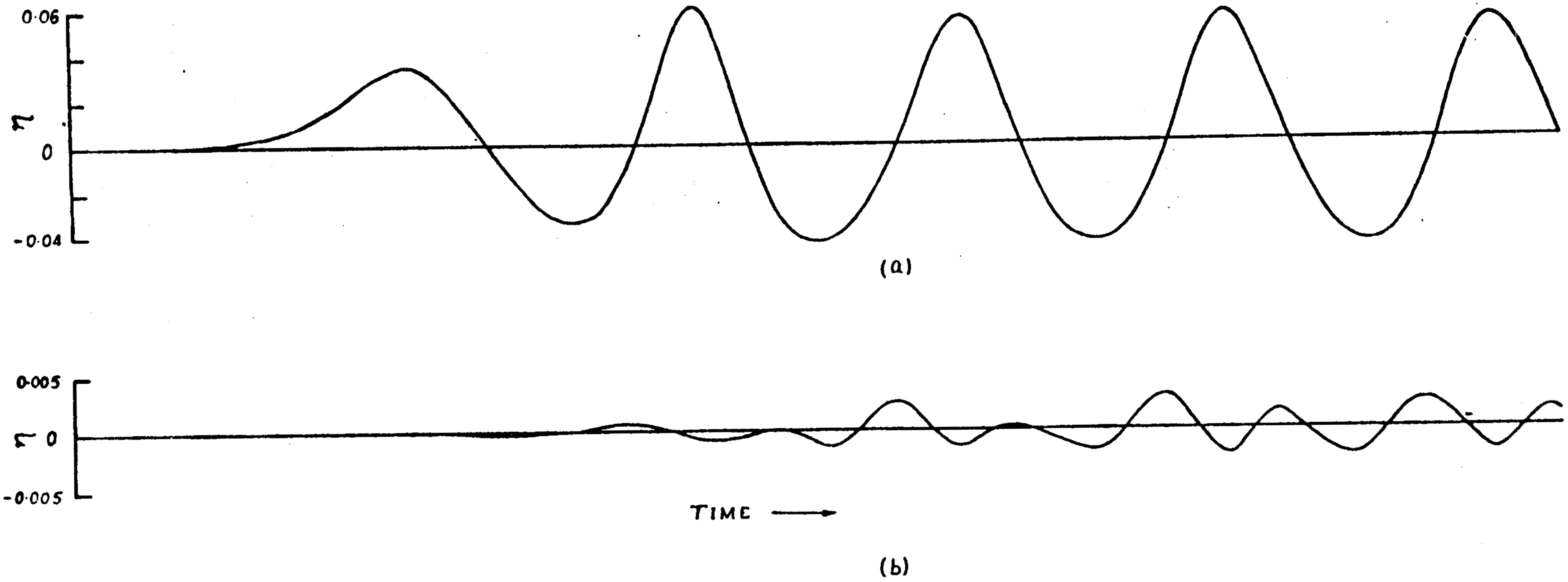

FIGURE 7·1

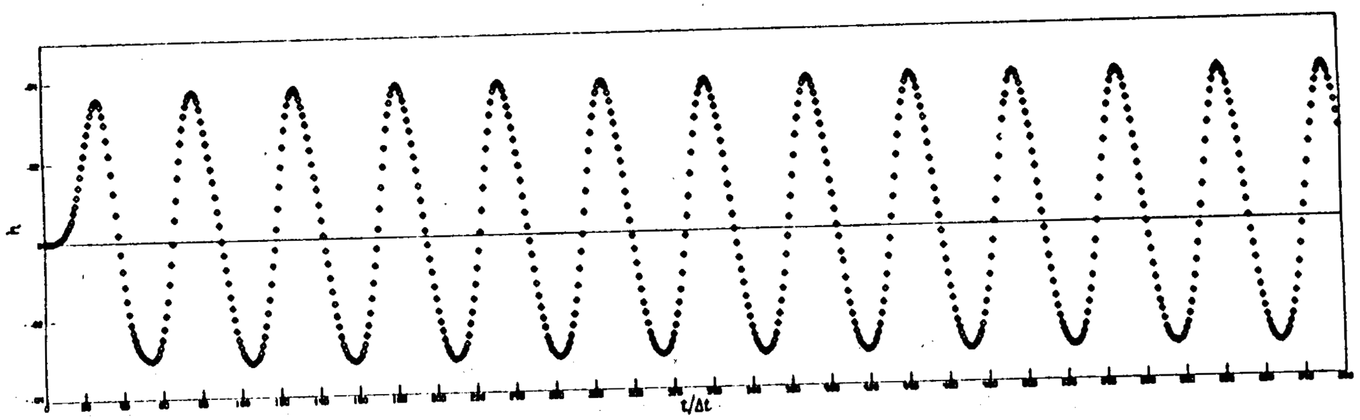

FIGURE 7.2

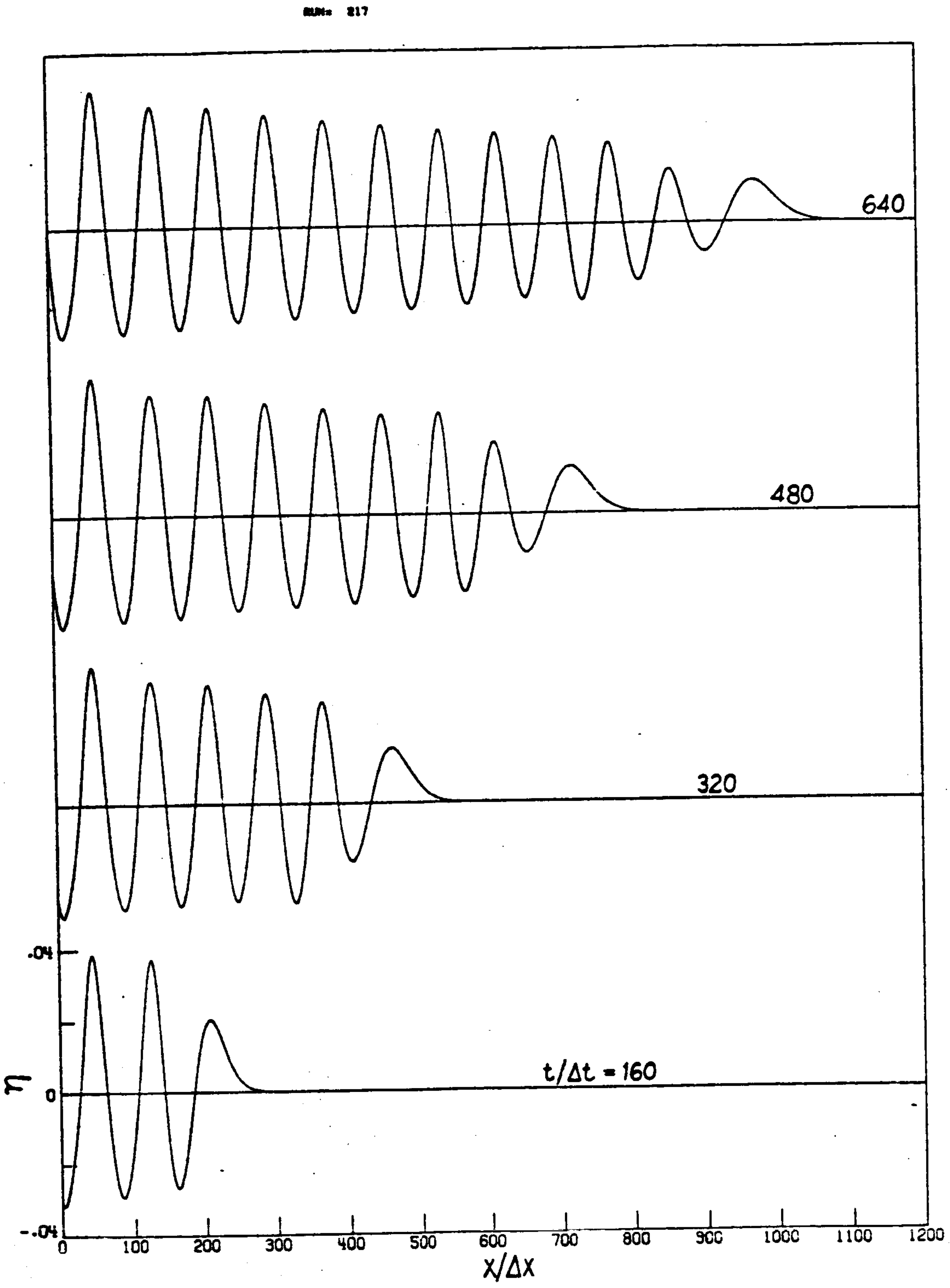
RUN= 217
640
480
320
t/Δt = 160
η
.04
0
-.04
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
X/ΔX

FIGURE 7.3 (a)

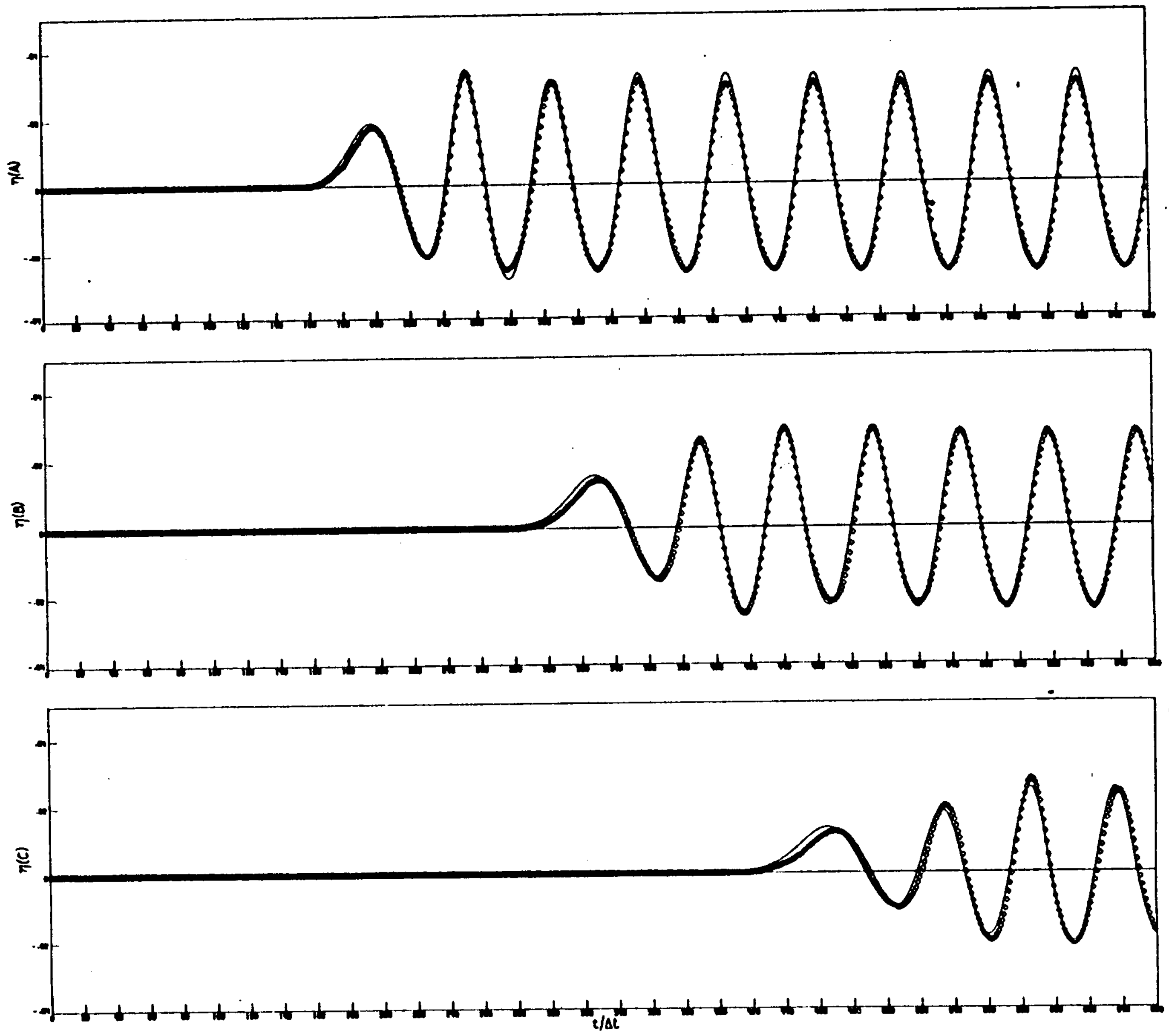

FIGURE 7.3 (b)

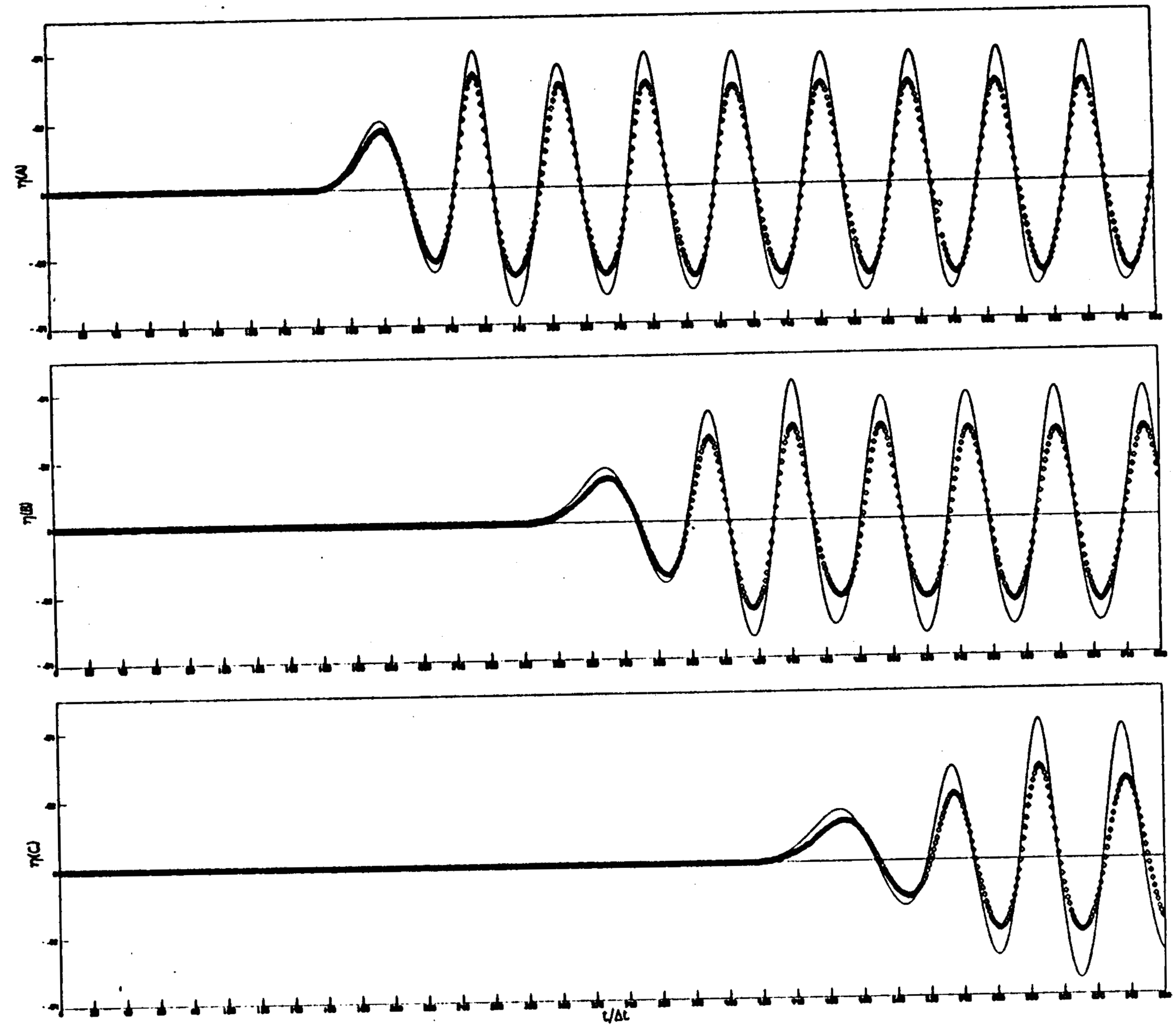

FIGURE 7.4

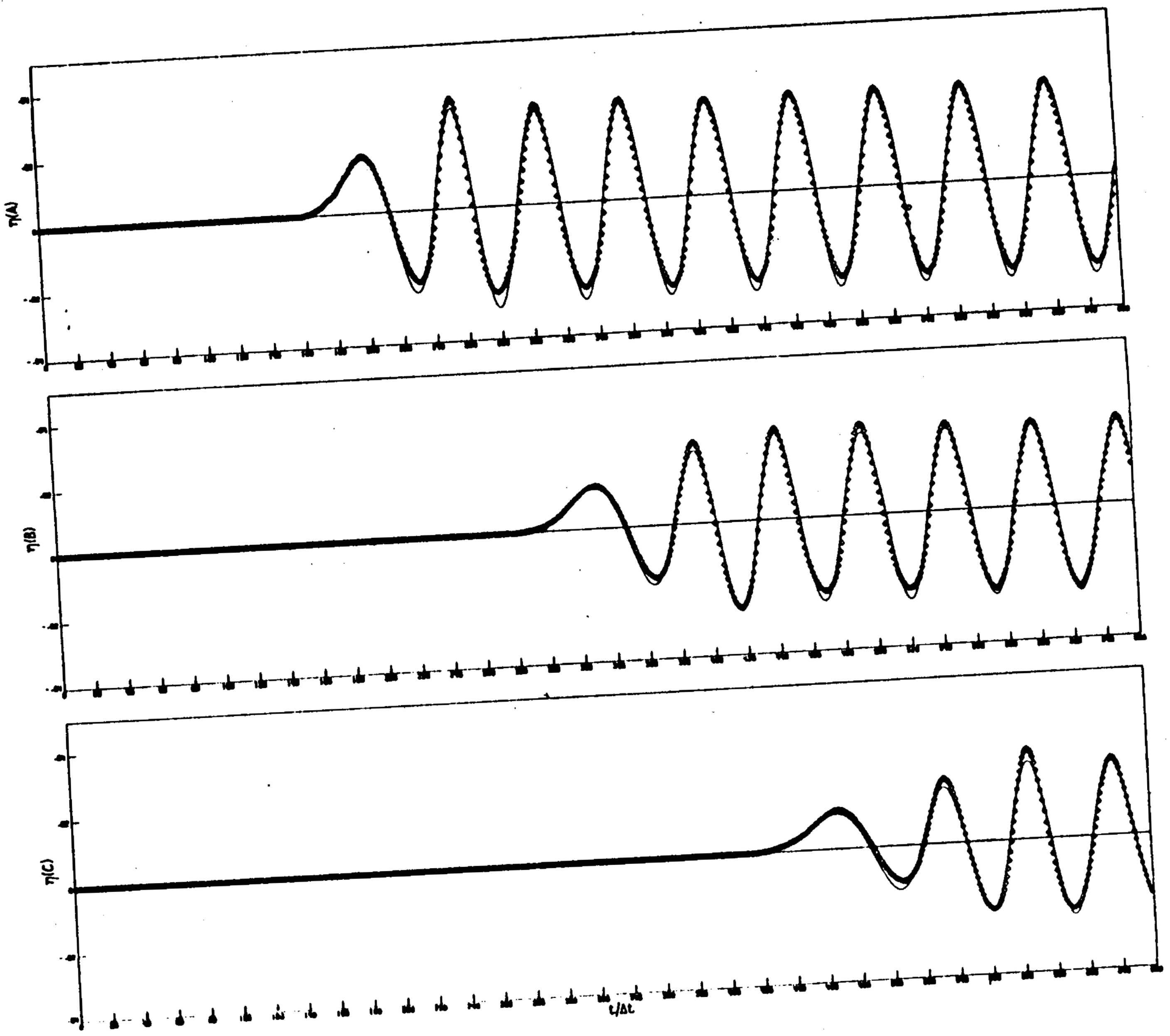

FIGURE 7.5

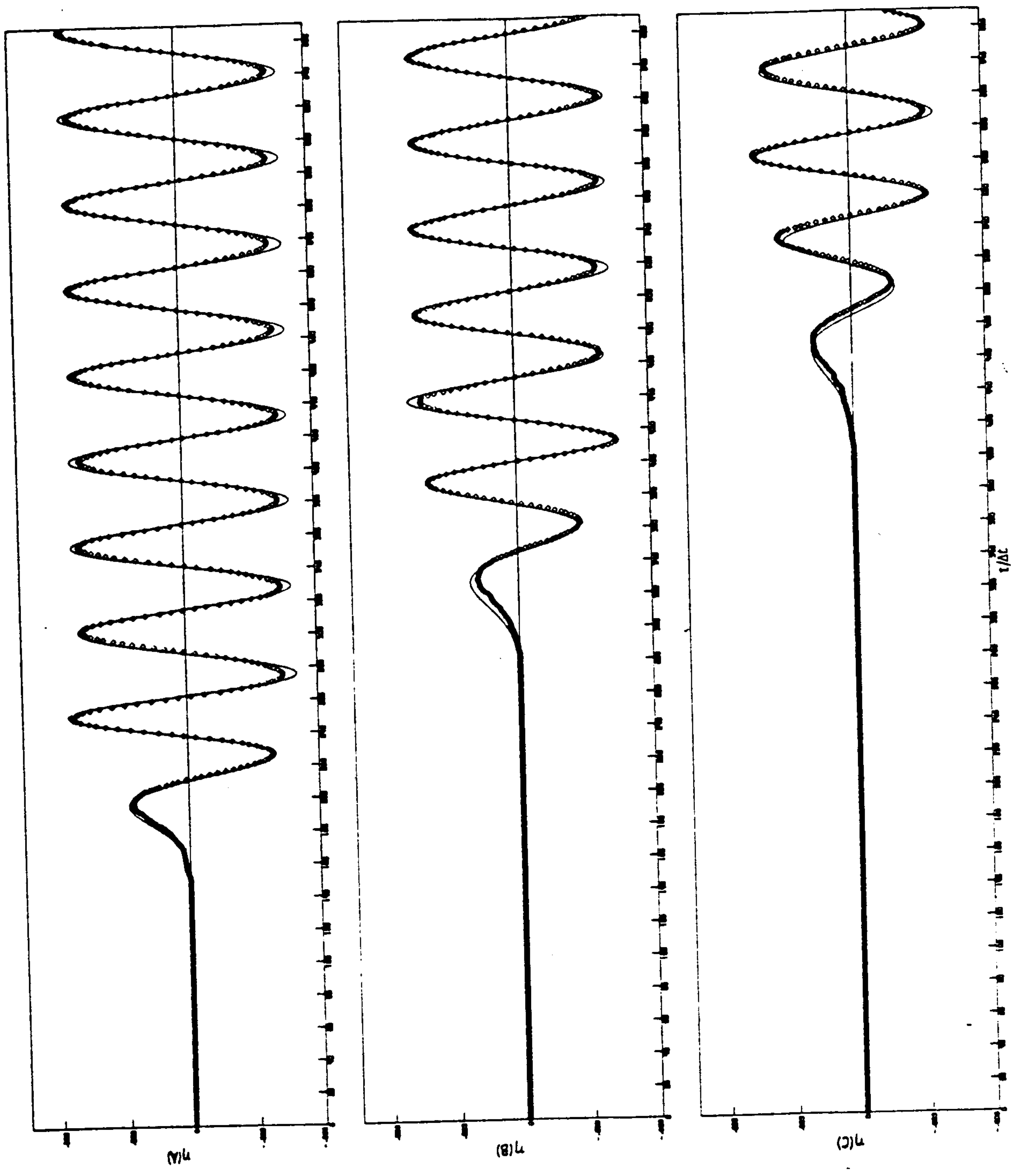

FIGURE 7.6

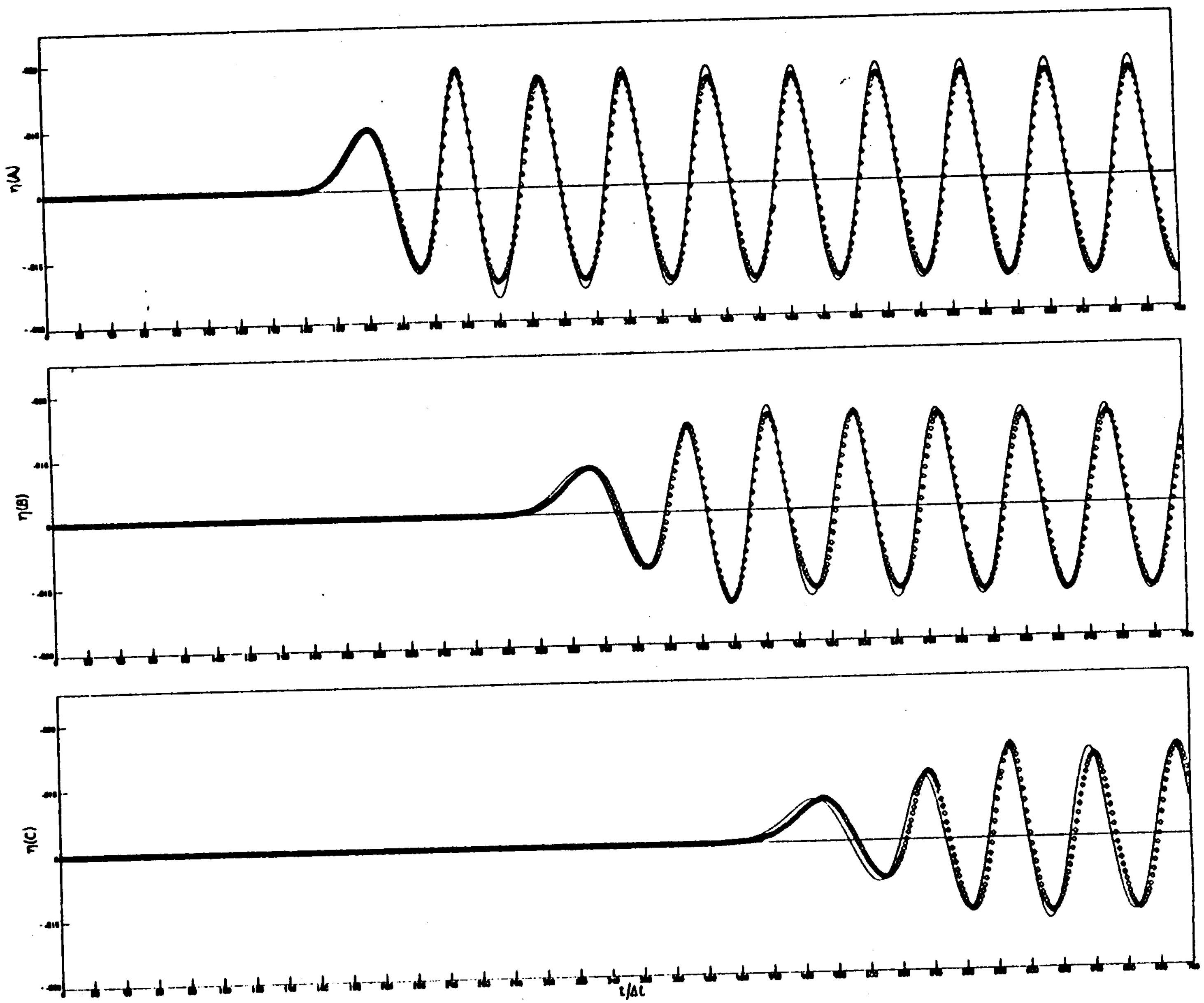

FIGURE 7.7

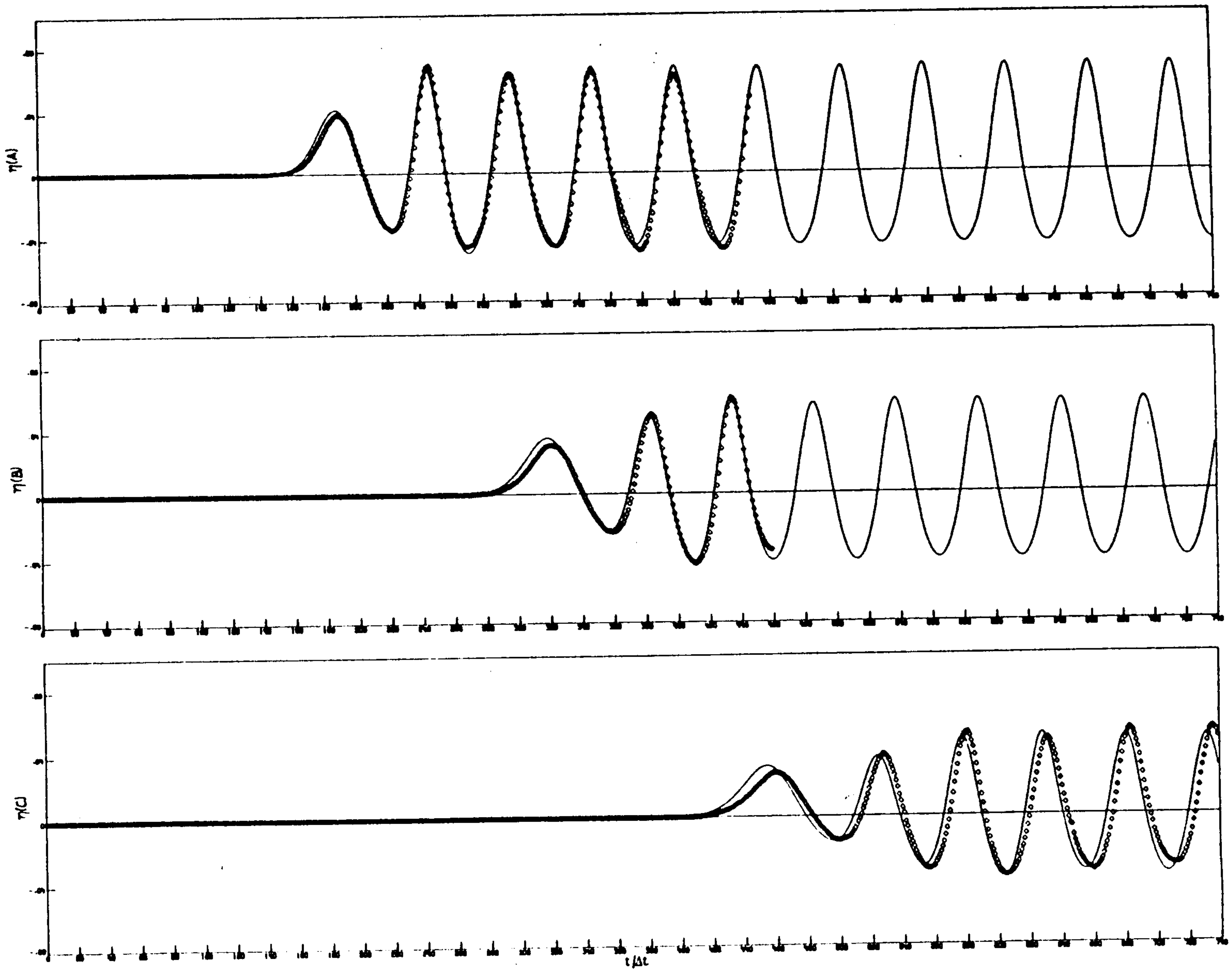

FIGURE 7.8

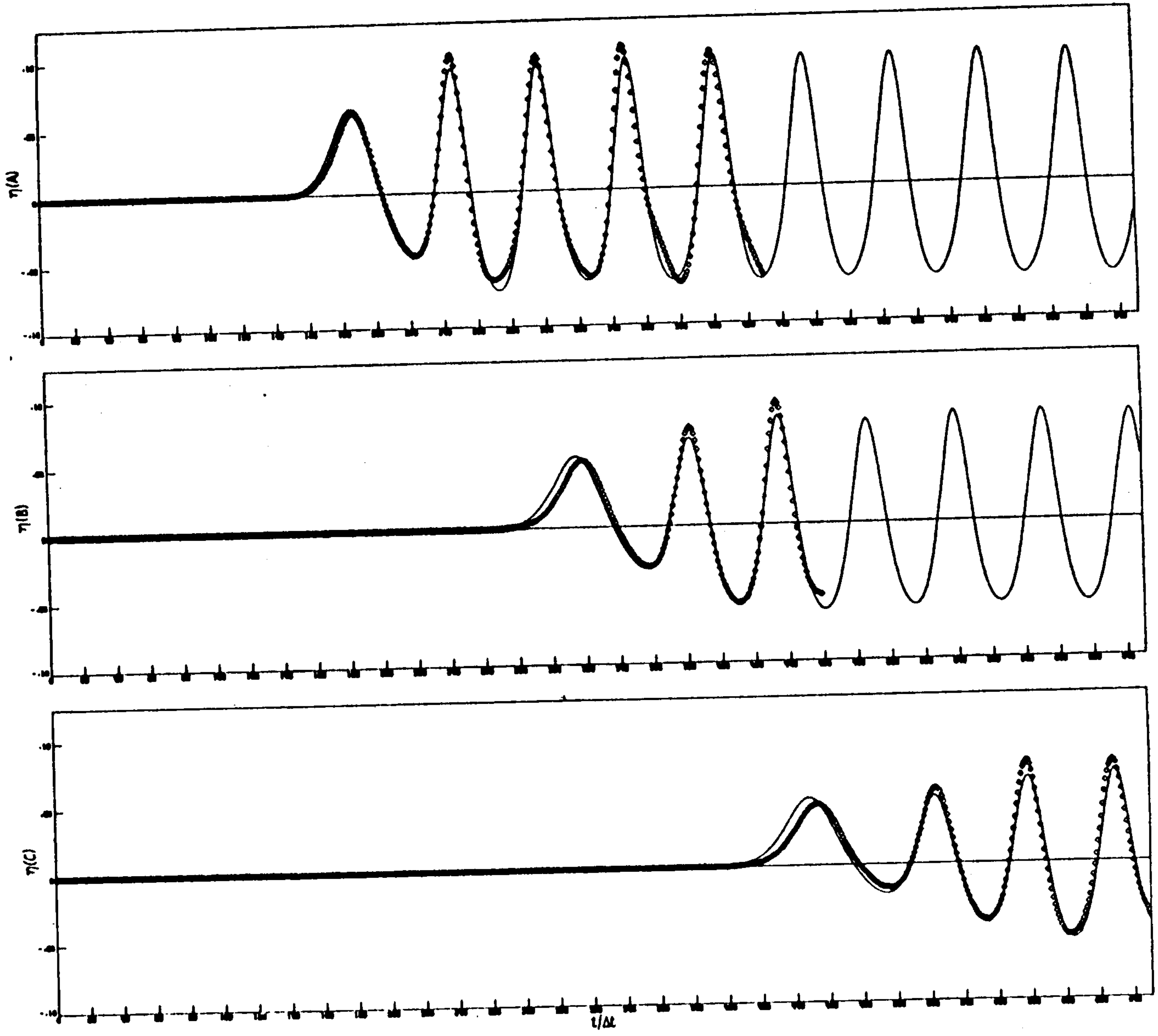

FIGURE 7.9

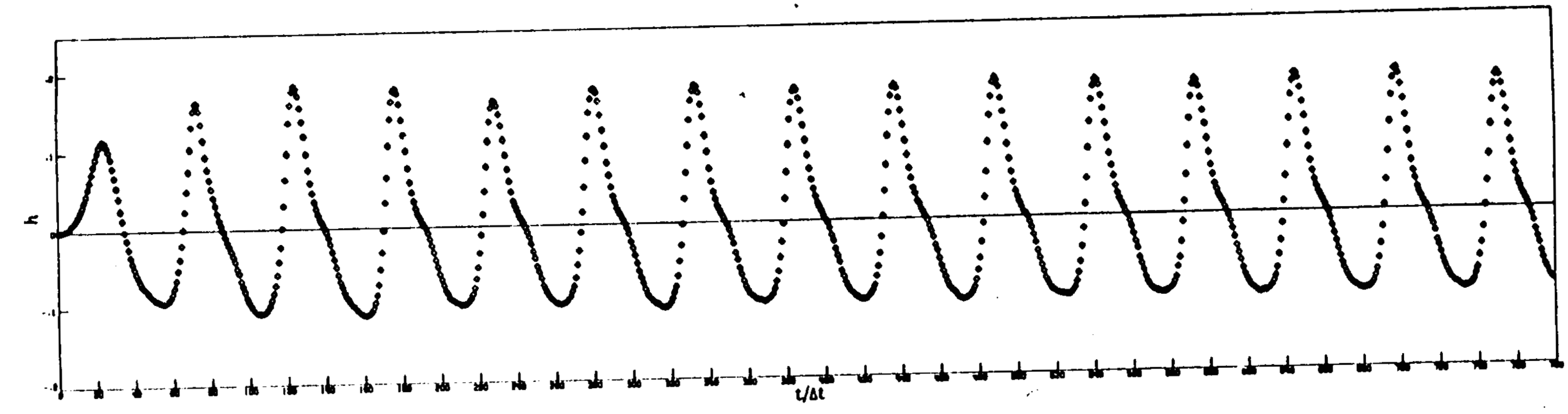

FIGURE 7.10

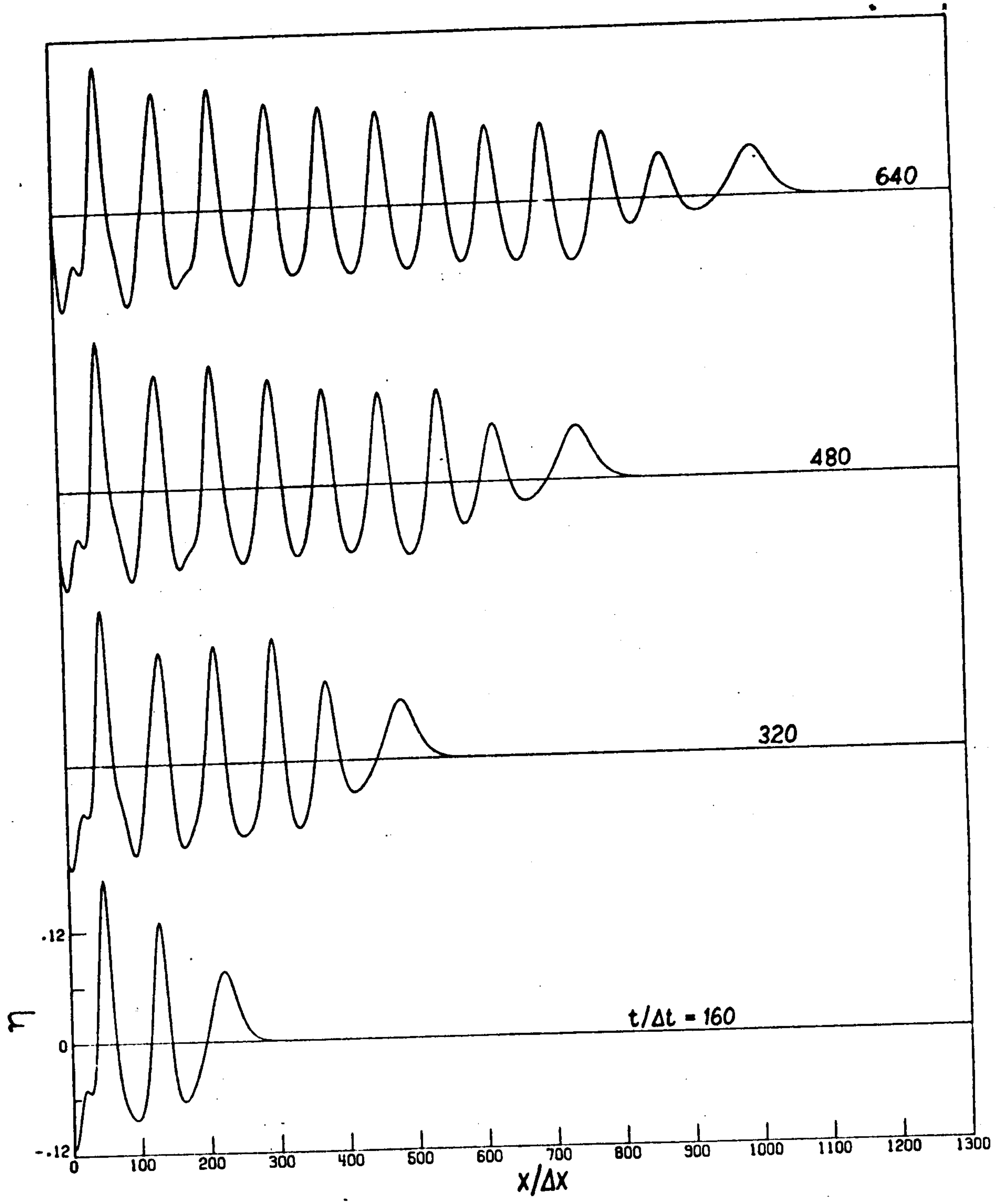
640
480
320
t/Δt = 160
η
.12
0
-.12
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
x/Δx

FIGURE 7.11(a)

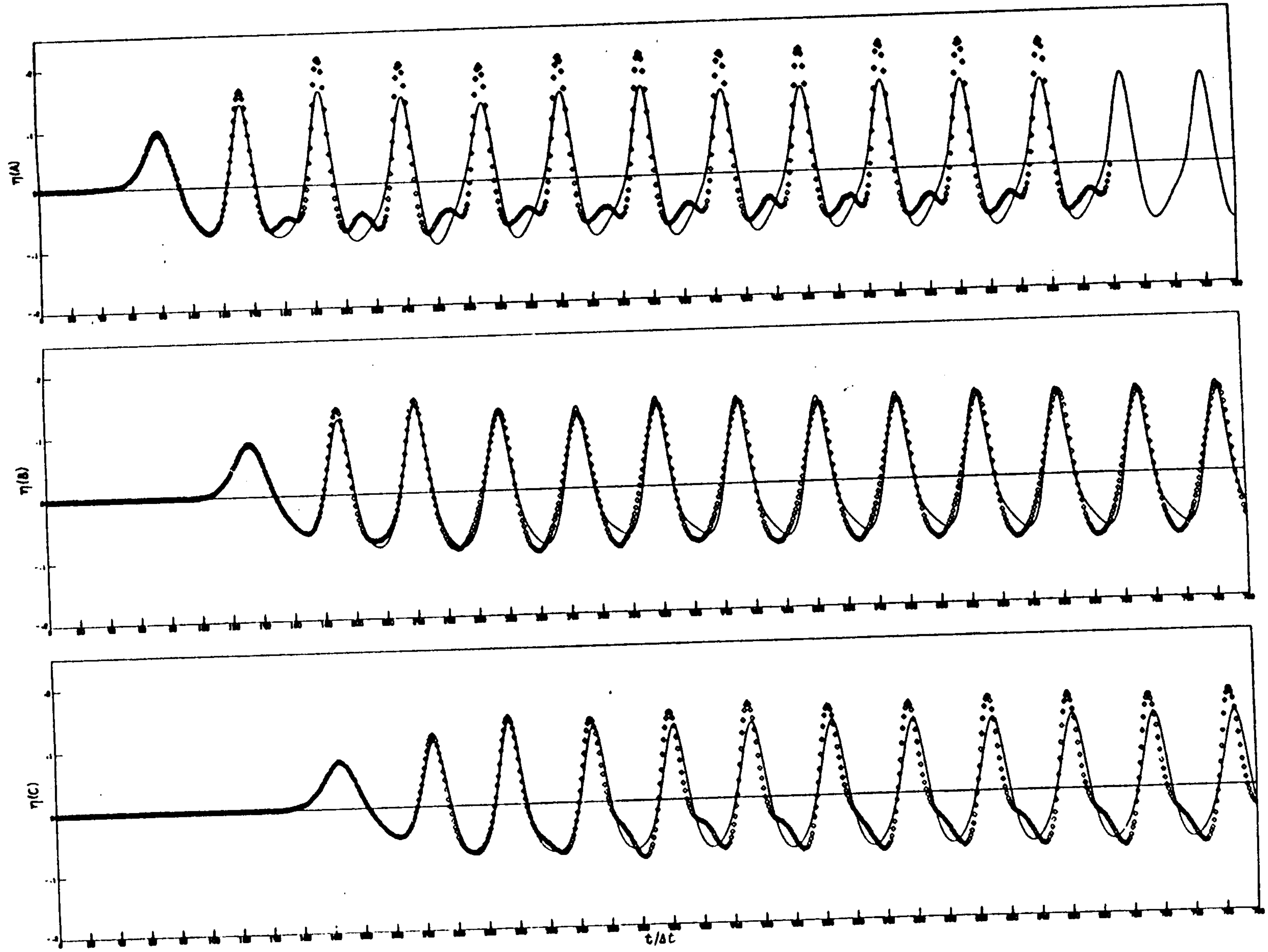

FIGURE 7.11(b)

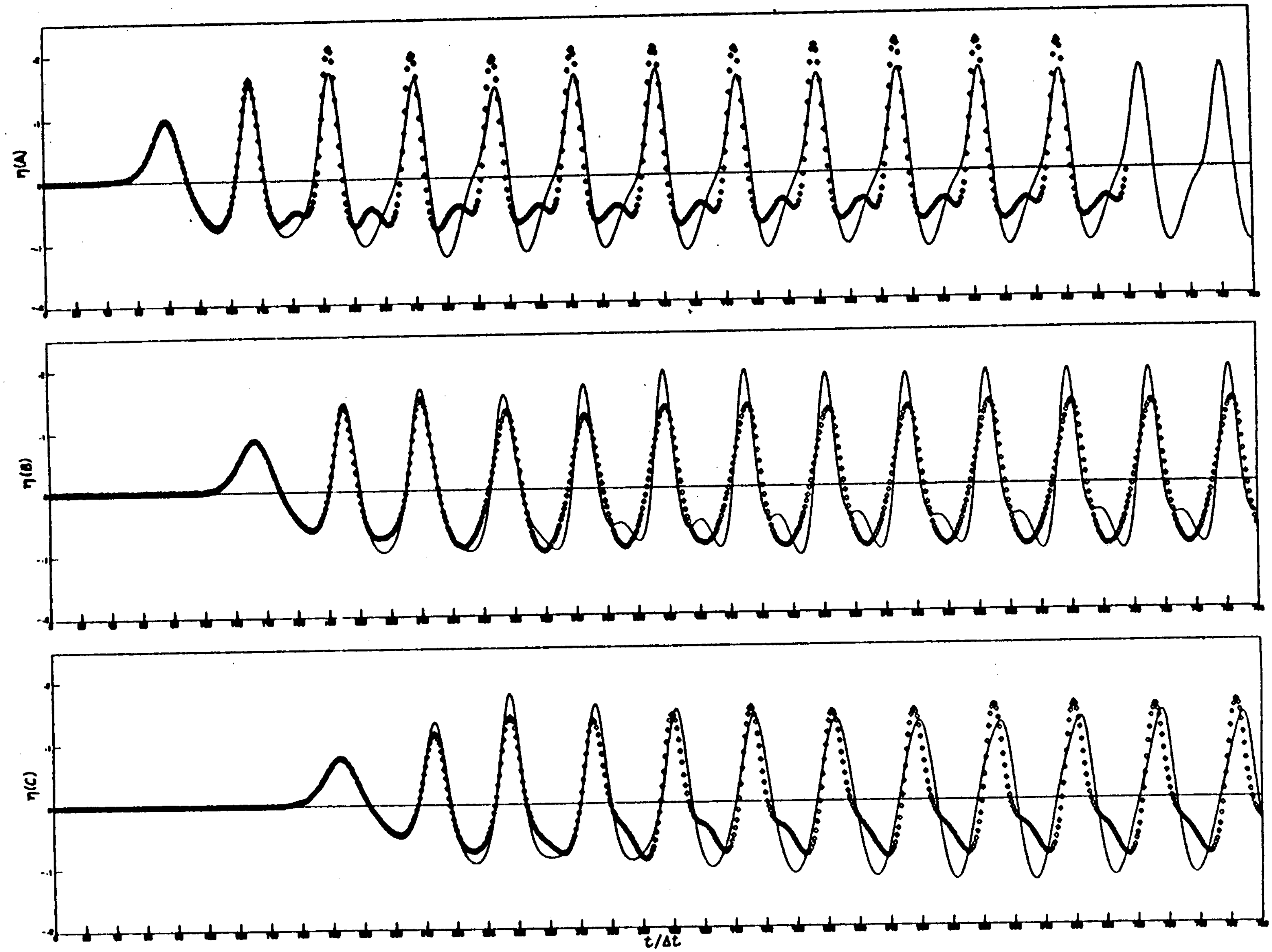

FIGURE 7.12

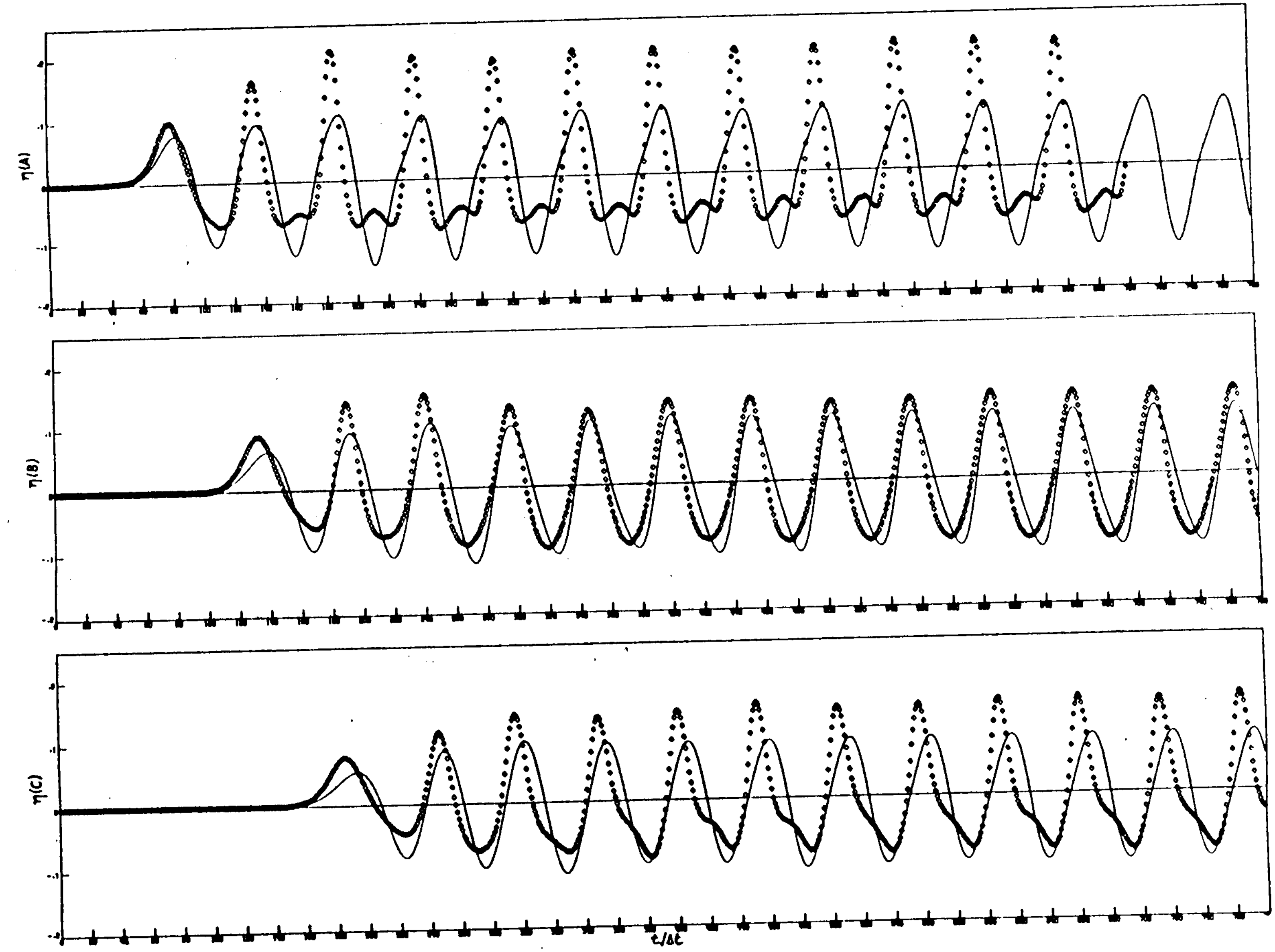

FIGURE 7.13

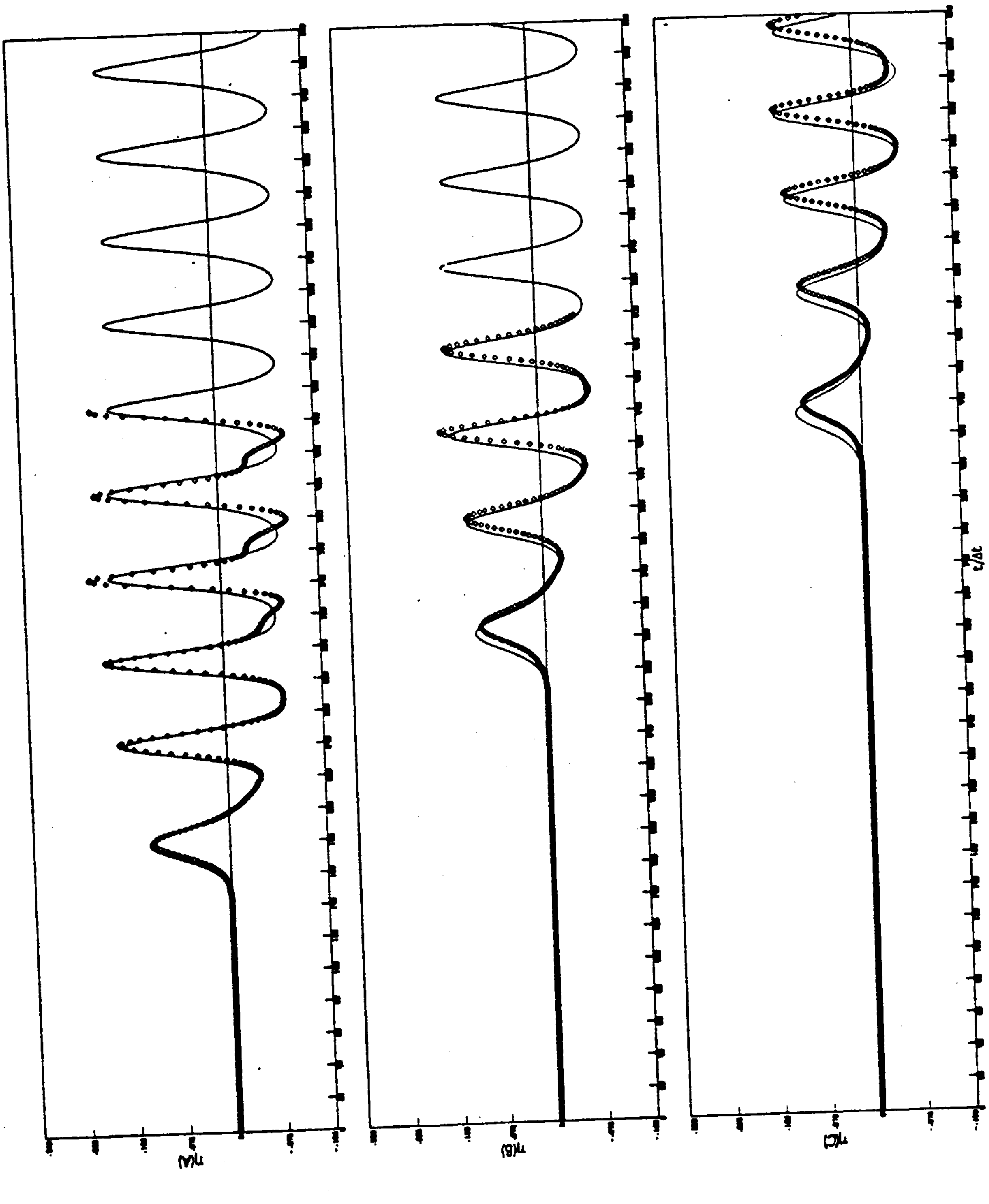

FIGURE 7.14

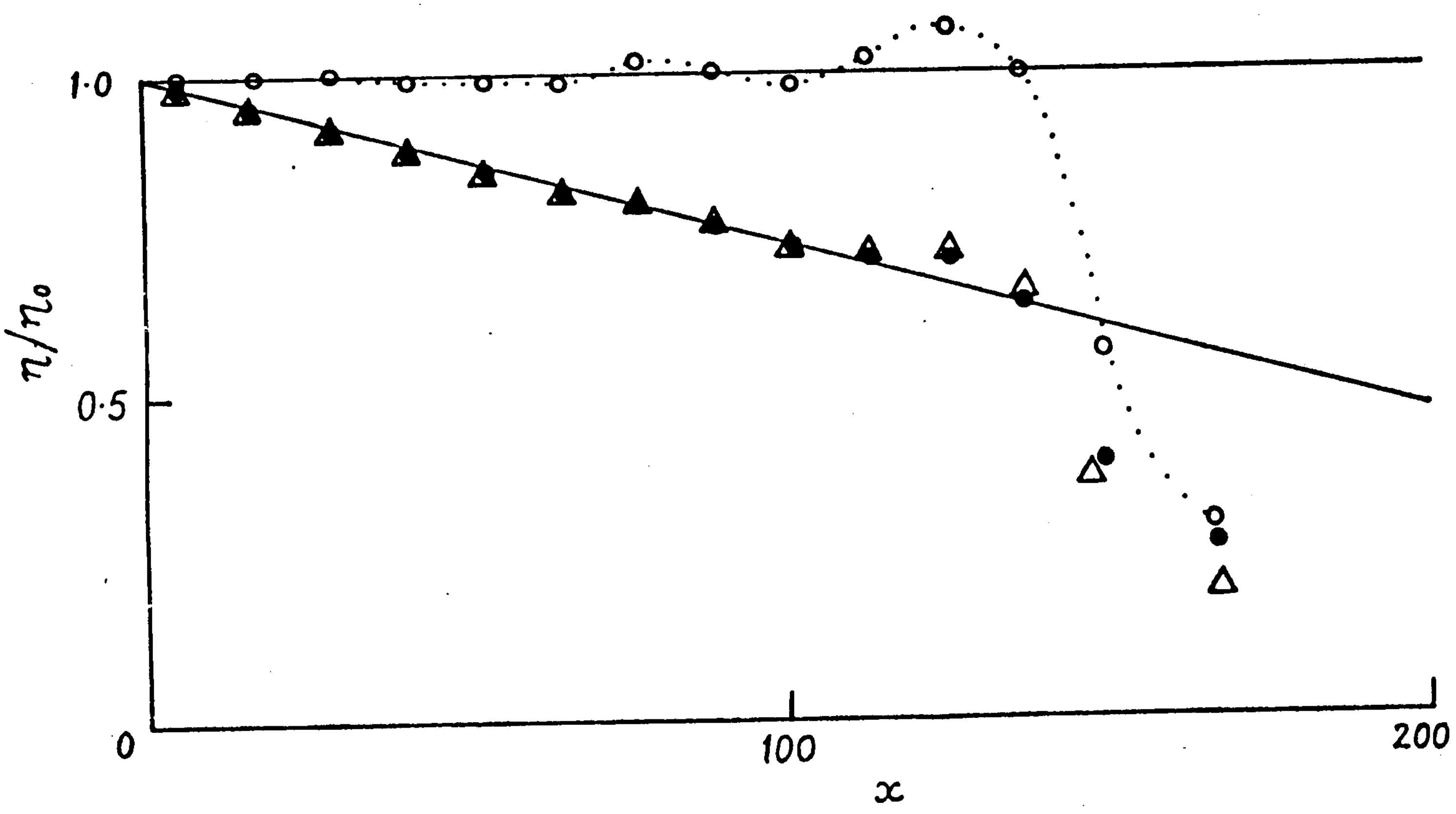

FIGURE 7·15

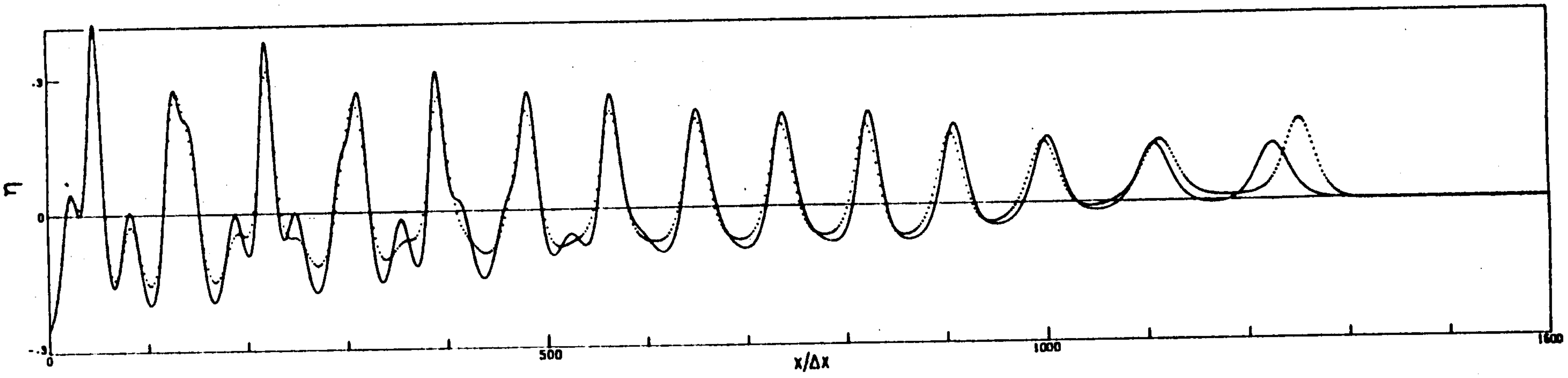

FIGURE 7.16

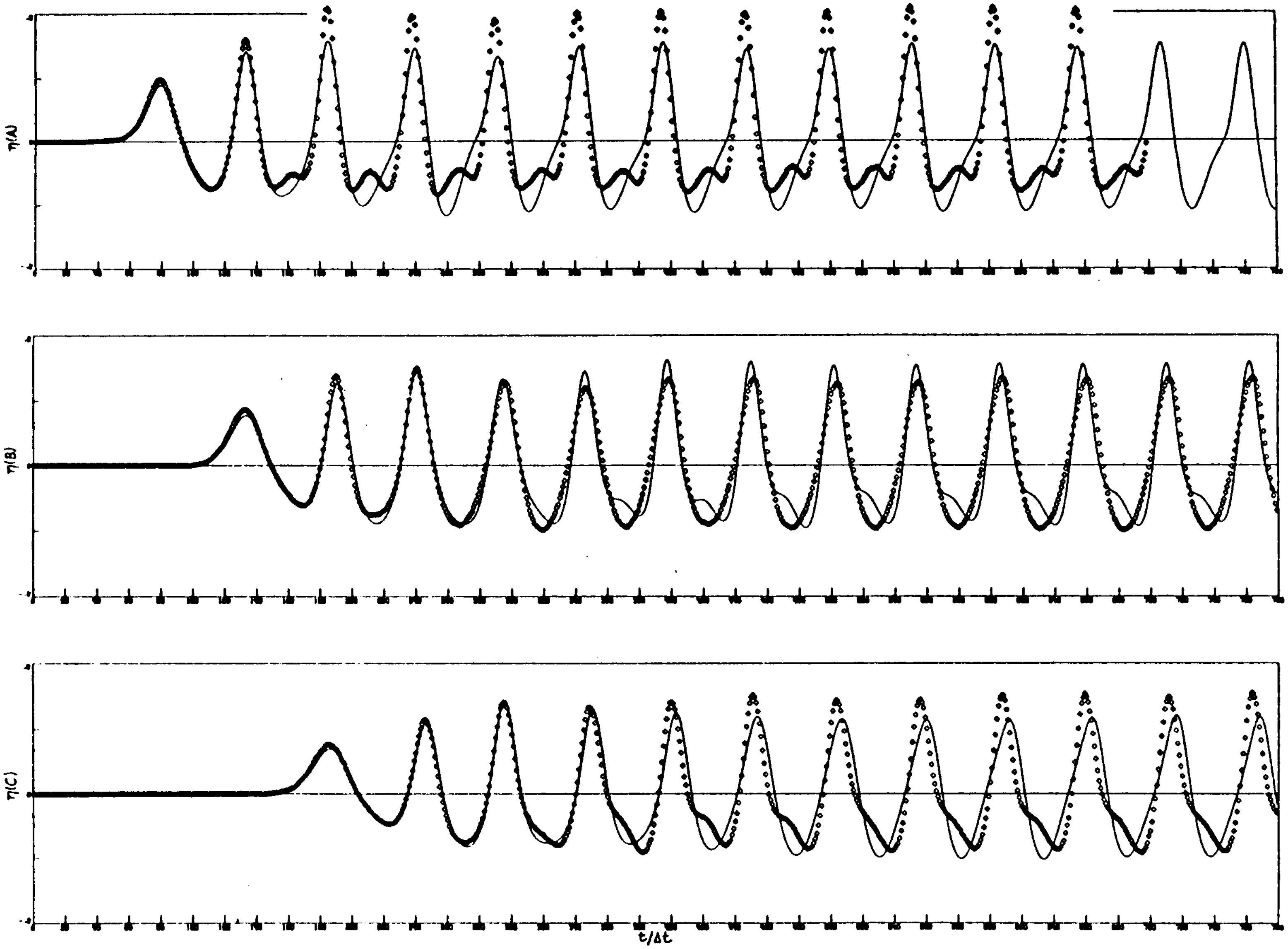

FIGURE 7.17

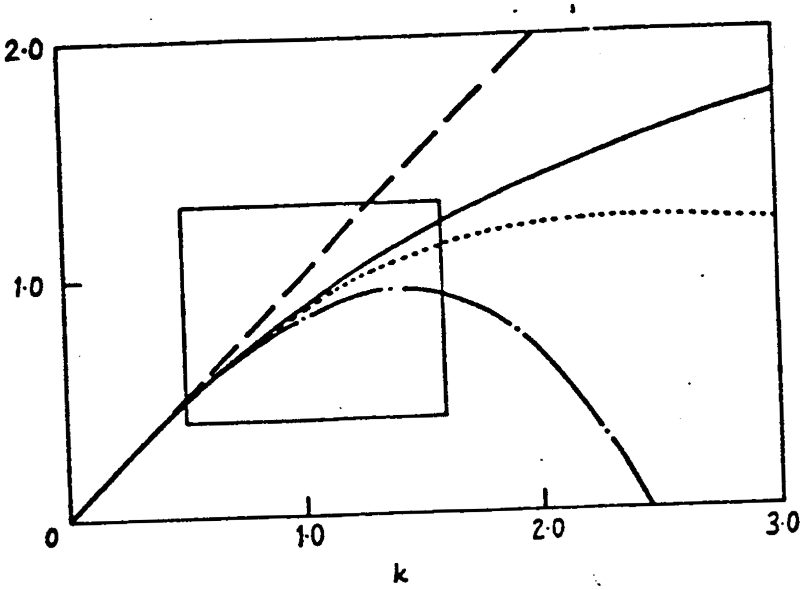

FIGURE 7·18 a

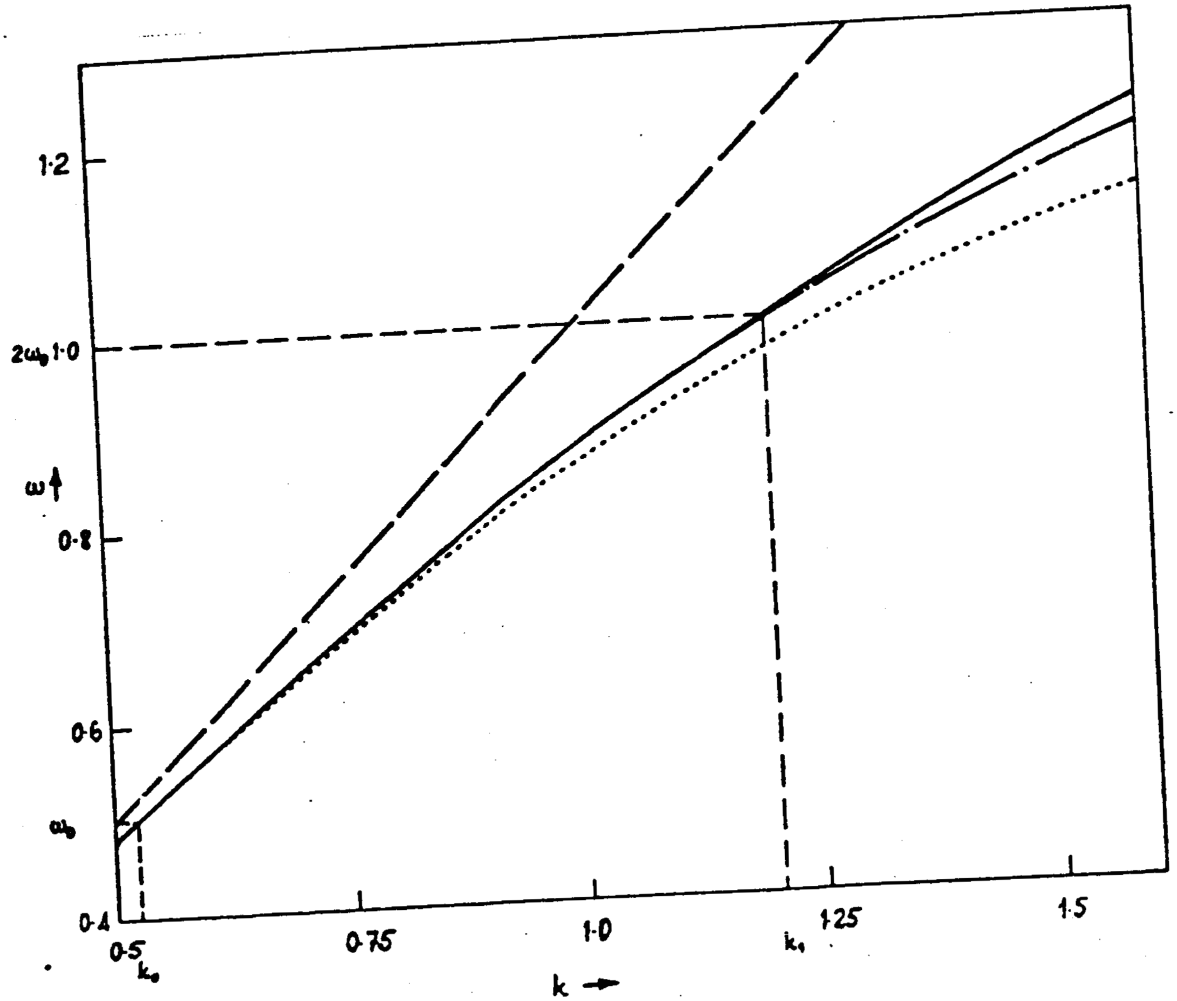

FIGURE 7·18 b

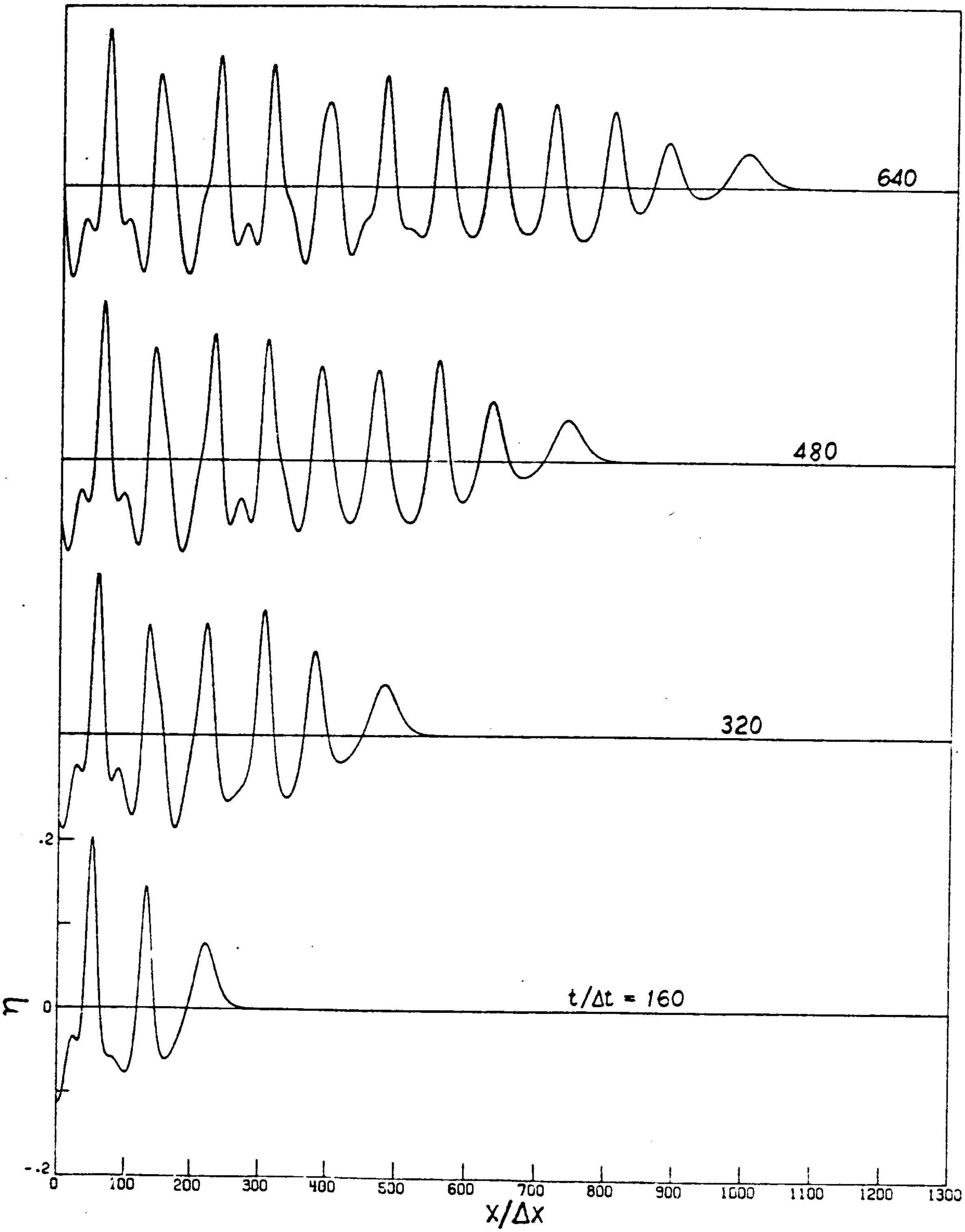
640
480
320
t/Δt = 160
η
.2
0
-.2
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
X/ΔX

FIGURE 7.19 (a)

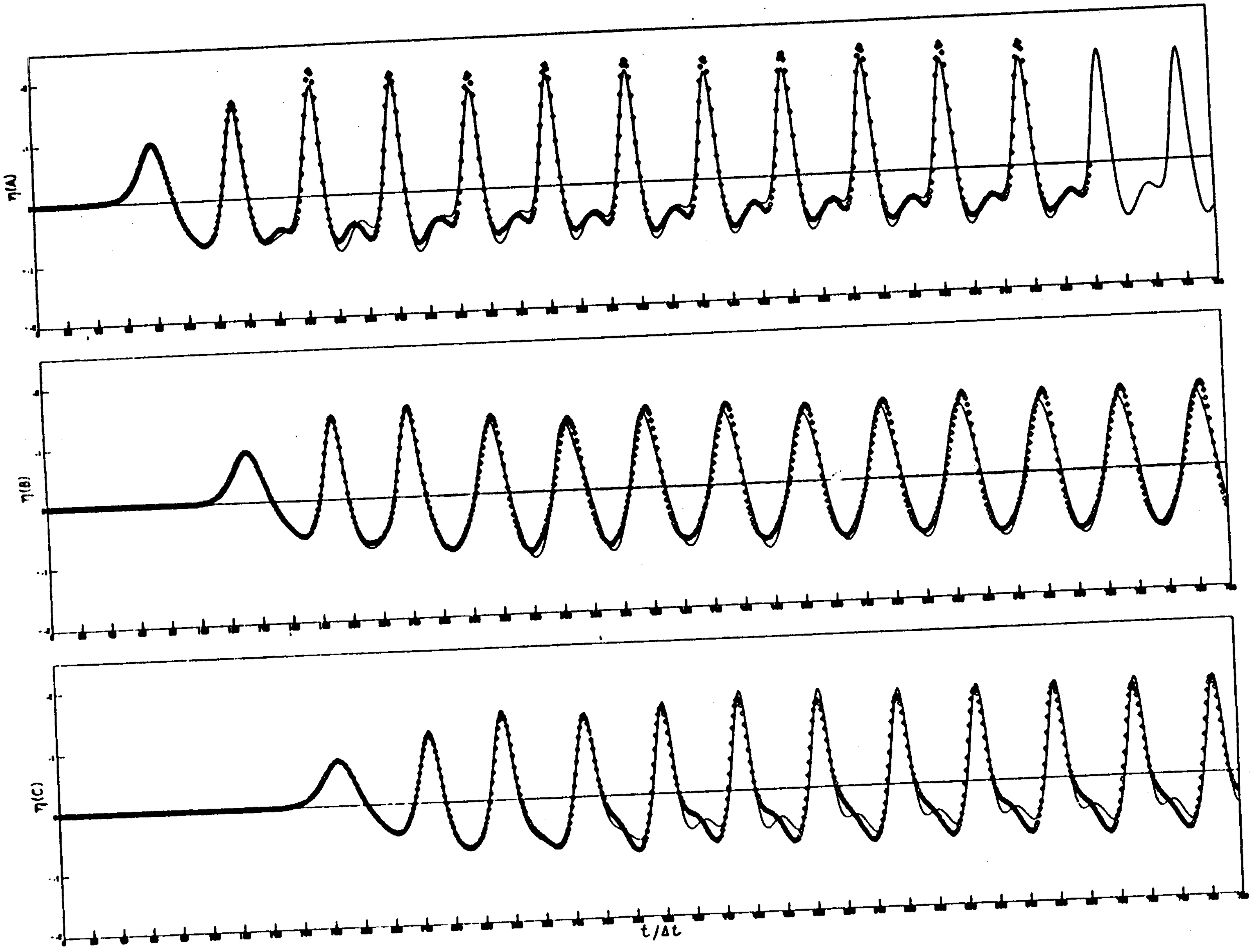

FIGURE 7.19(b)

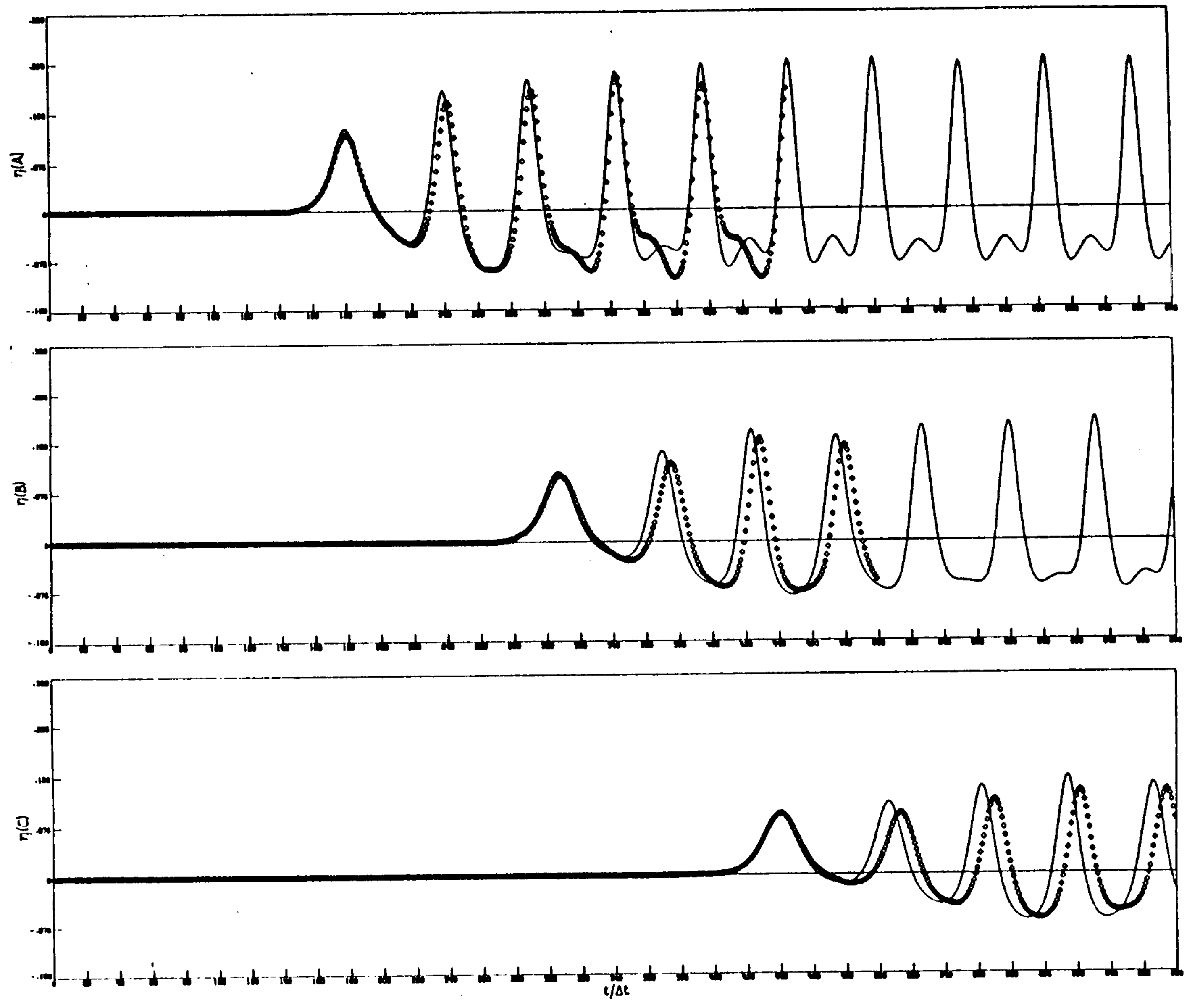

FIGURE 7.20

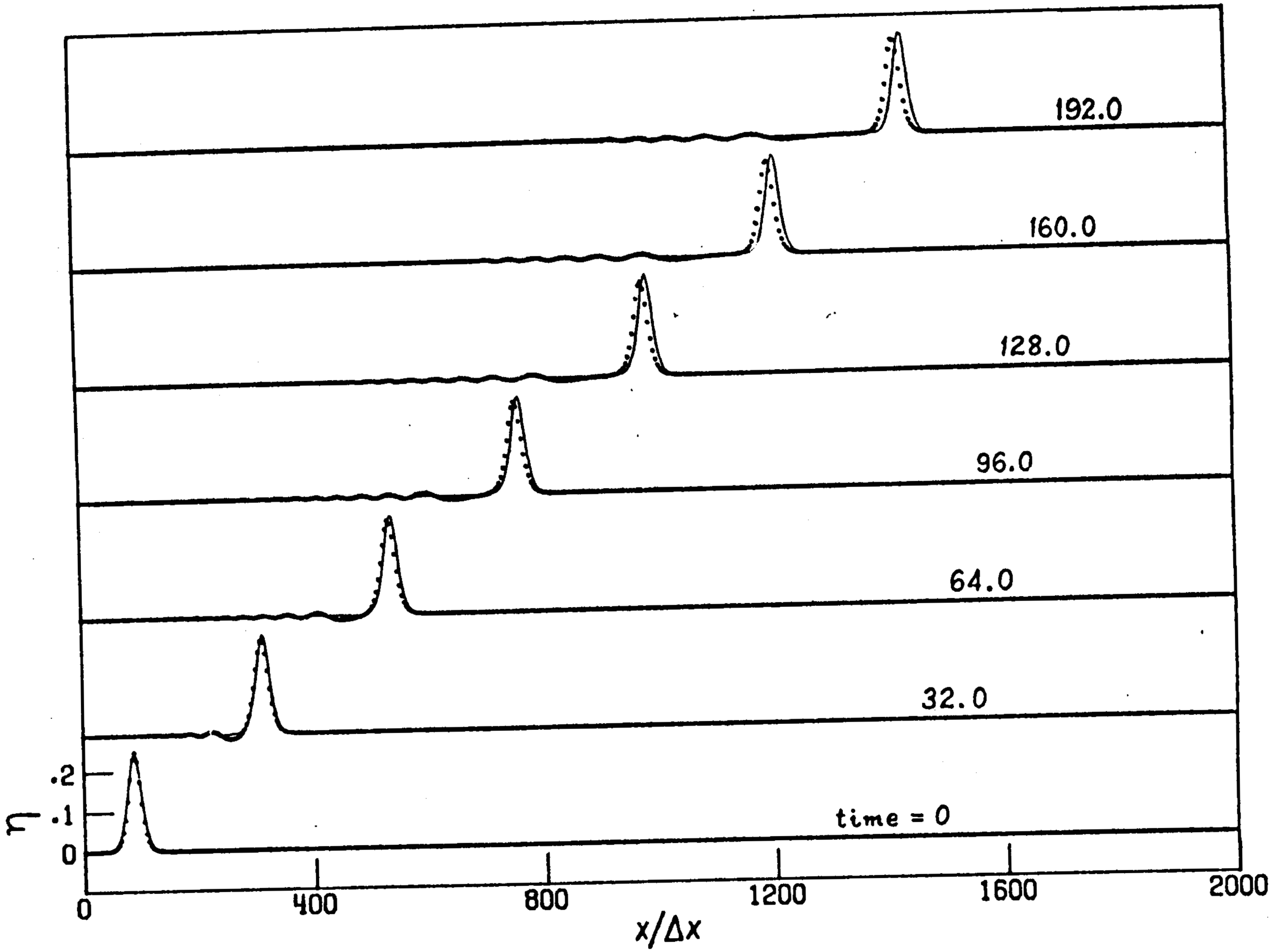
192.0
160.0
128.0
96.0
64.0
32.0
time = 0
η
.2
.1
0
0
400
800
1200
1600
2000
$x/\Delta x$

FIGURE A1

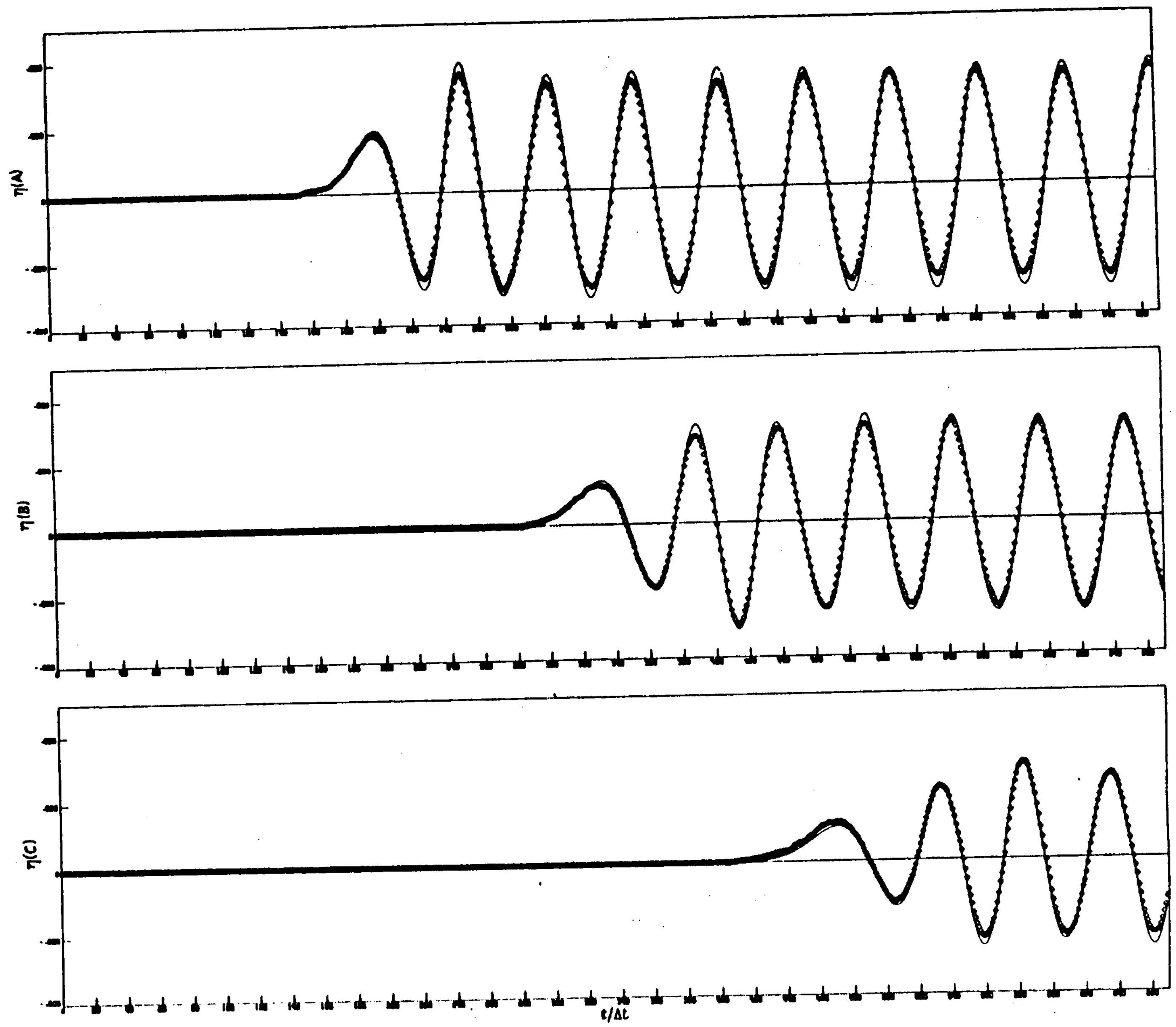

FIGURE B1

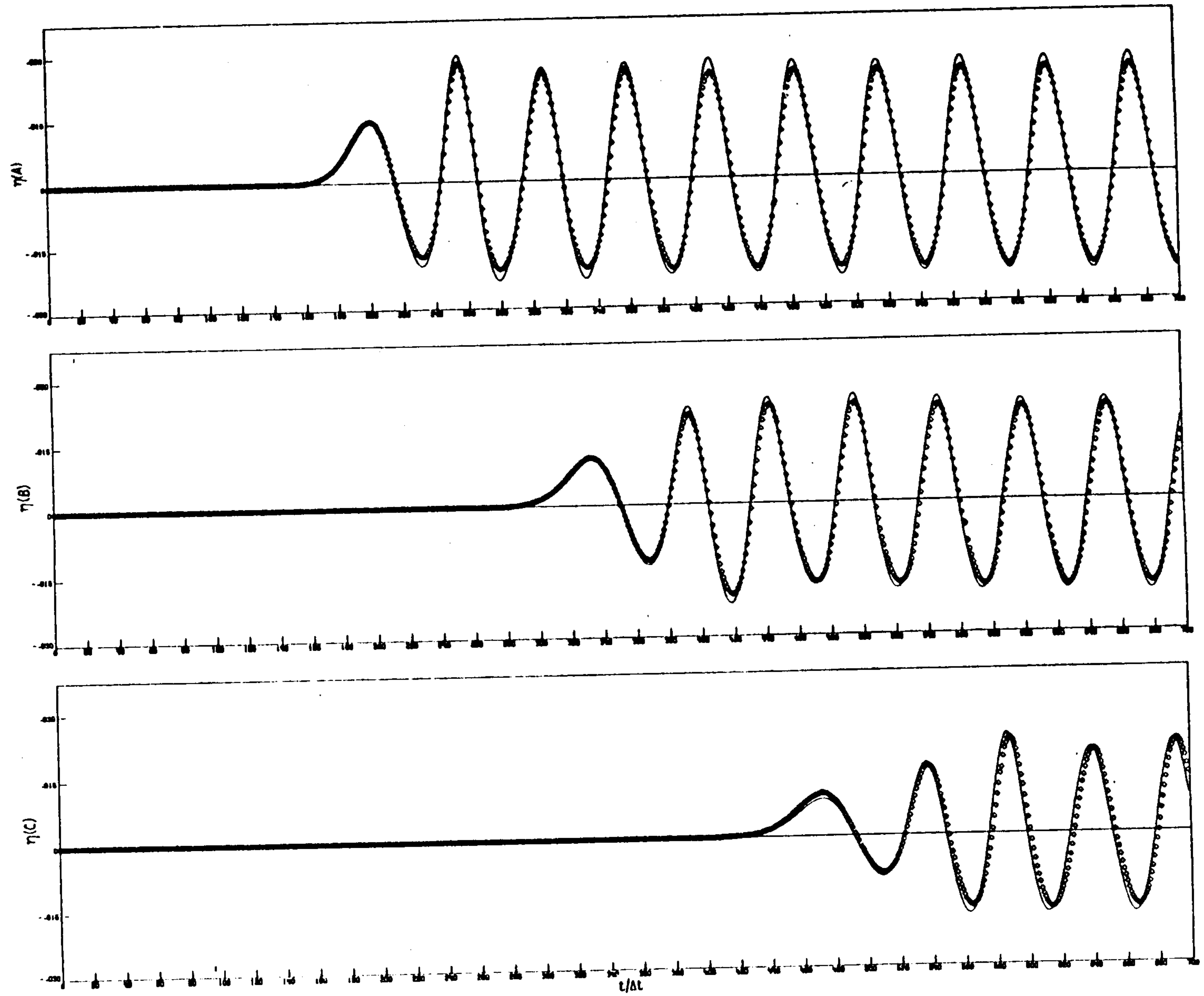

FIGURE B2

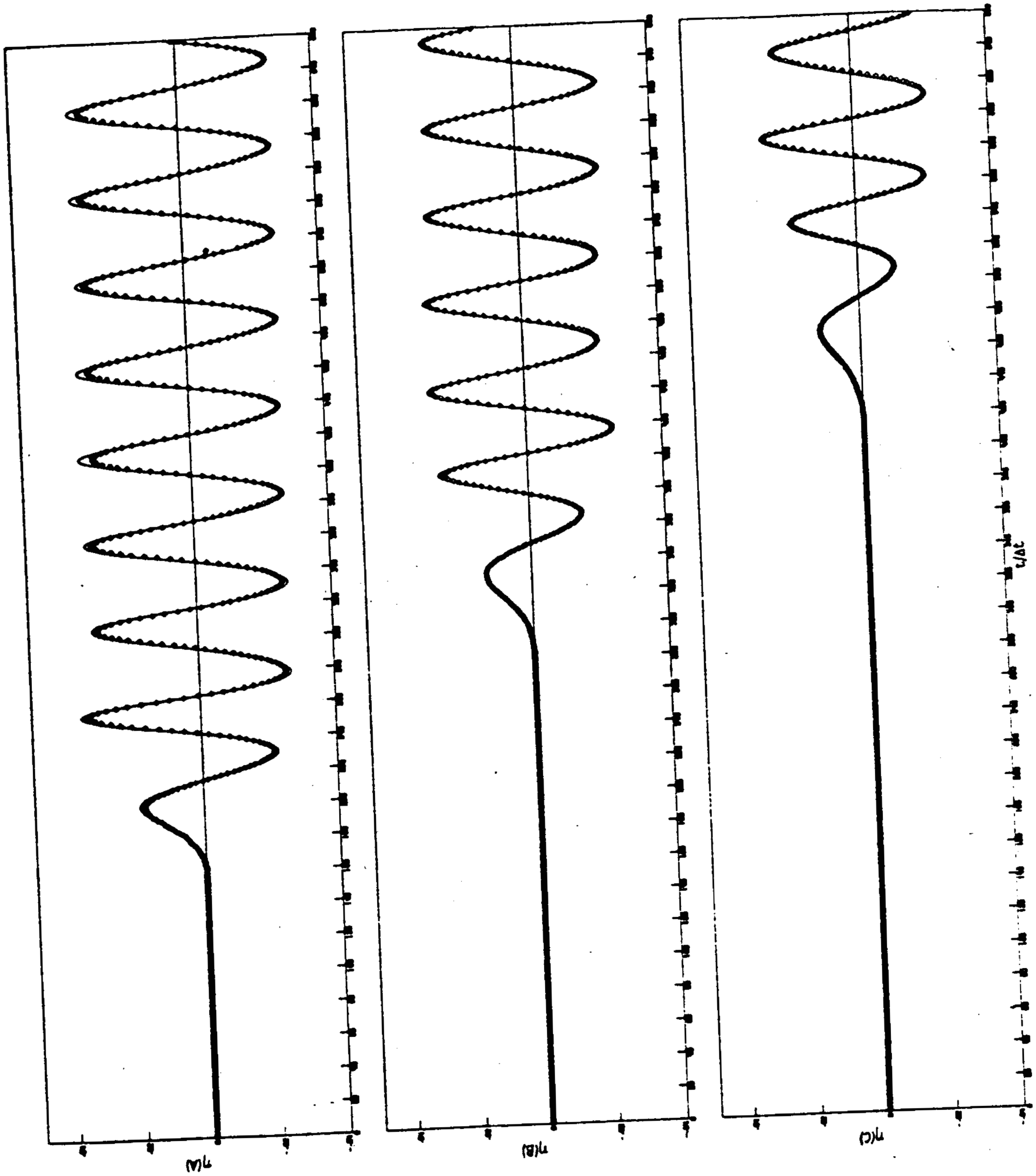

FIGURE B3

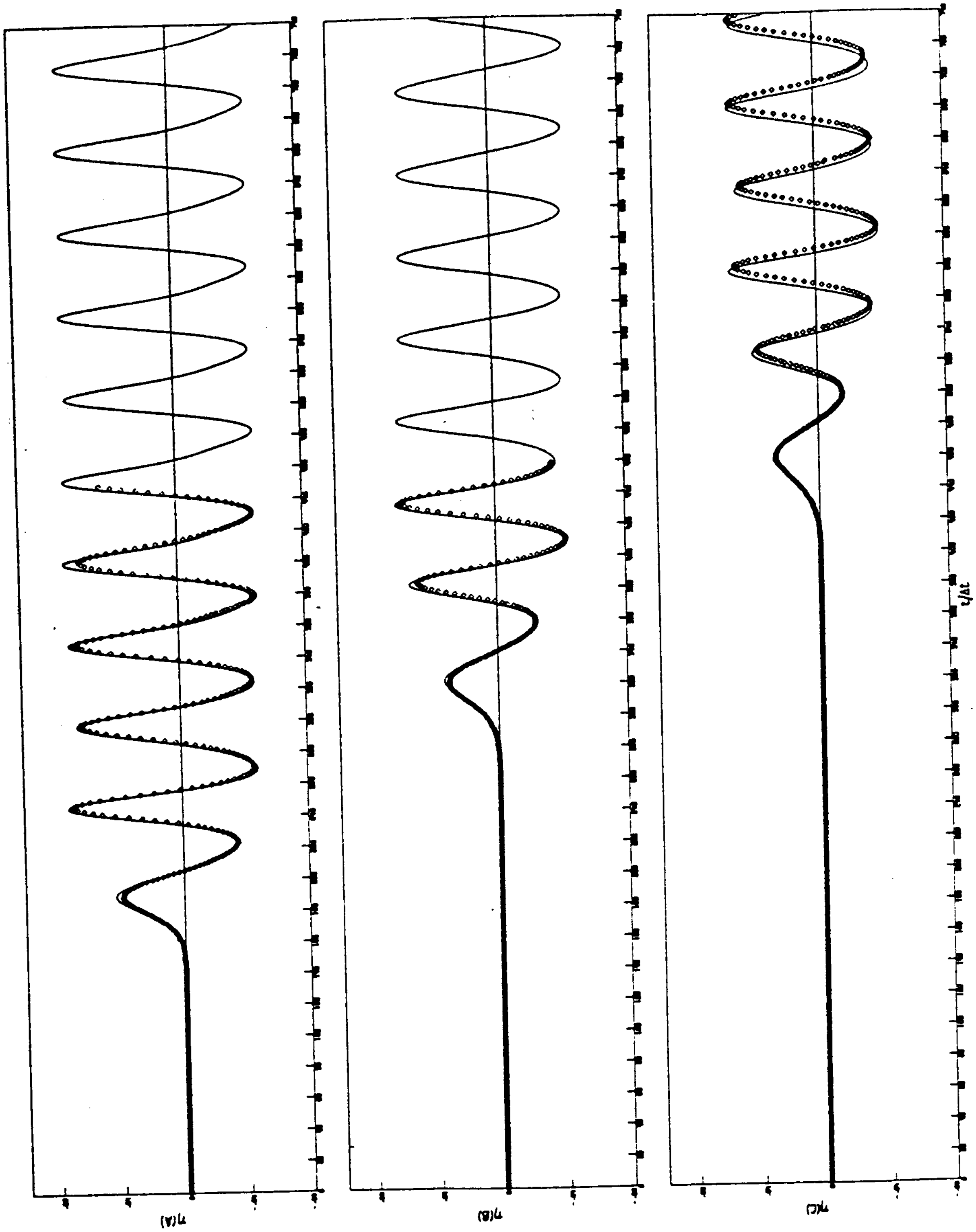

FIGURE B4

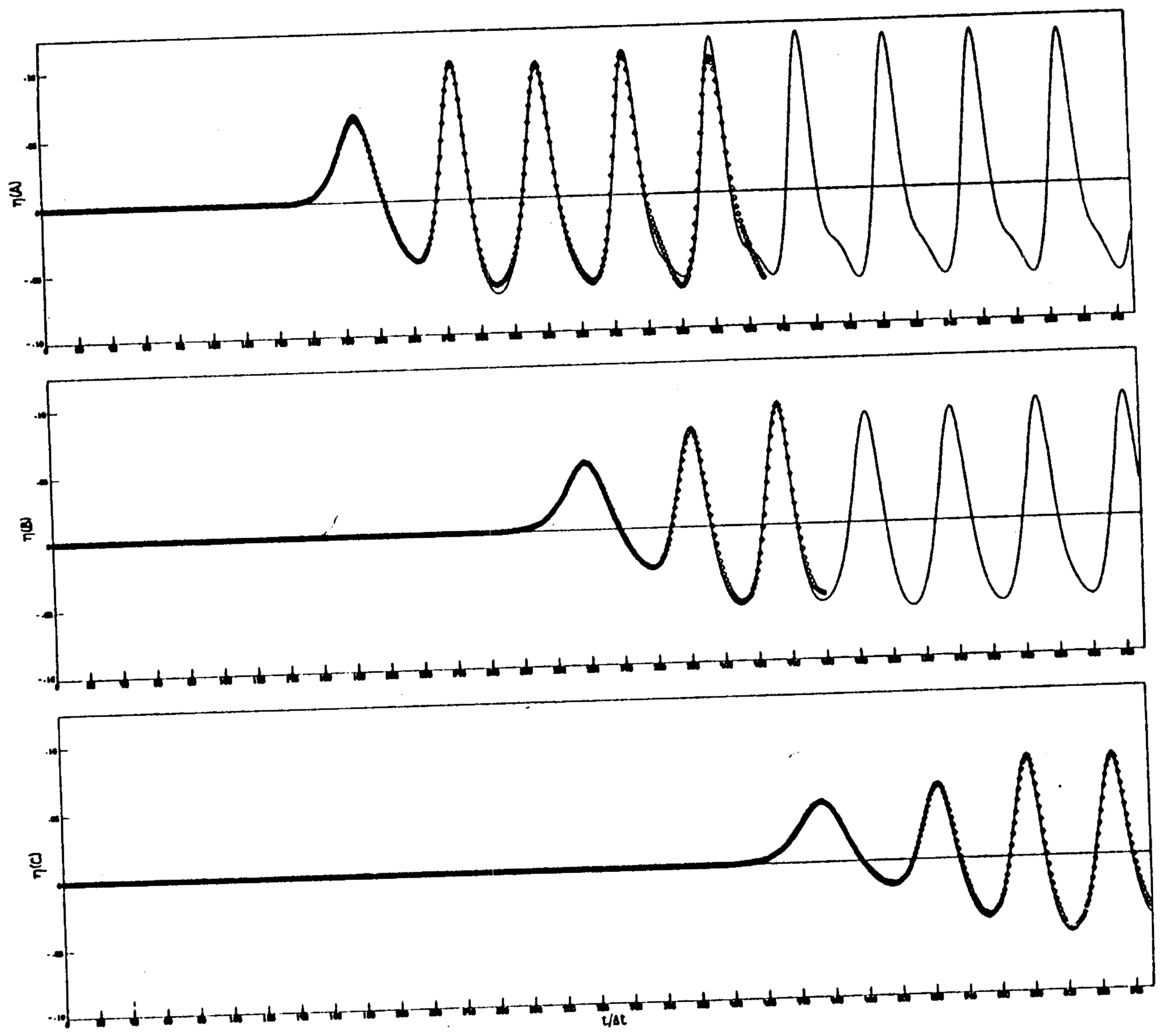

FIGURE B5

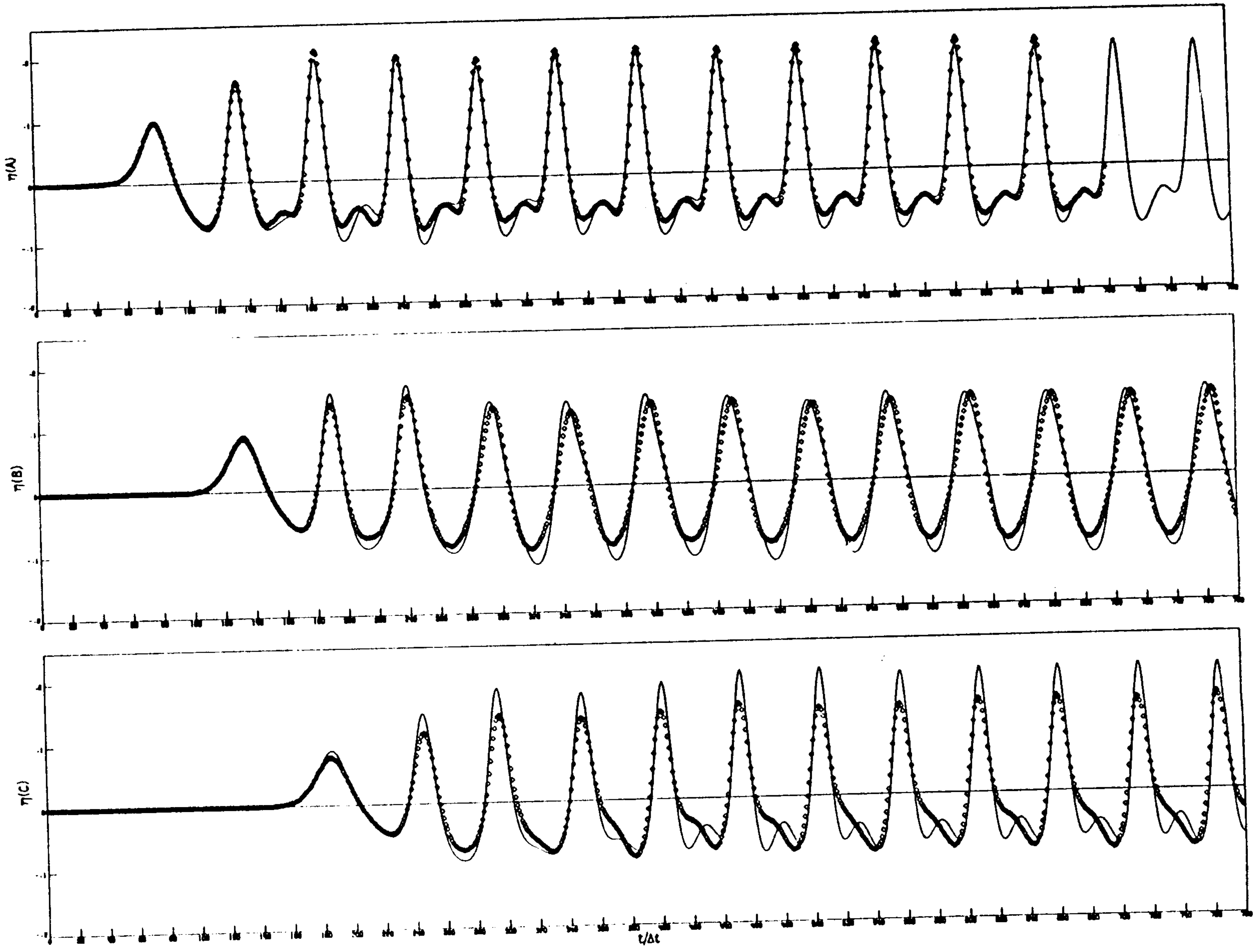

FIGURE B6